

Lingsong He
Bo Feng

Fundamentals of Measurement and Signal Analysis

 华中科技大学出版社
Huazhong University of Science and Technology Press

 Springer

Fundamentals of Measurement and Signal Analysis

Lingsong He · Bo Feng

Fundamentals of Measurement and Signal Analysis

Lingsong He
School of Mechanical Science
and Engineering
Huazhong University of Science
and Technology
Wuhan, Hubei, China

Bo Feng
School of Mechanical Science
and Engineering
Huazhong University of Science
and Technology
Wuhan, Hubei, China

ISBN 978-981-19-6548-7 ISBN 978-981-19-6549-4 (eBook)
<https://doi.org/10.1007/978-981-19-6549-4>

Jointly published with Huazhong University of Science and Technology Press
The print edition is not for sale in China (Mainland). Customers from China (Mainland) please order the
print book from: Huazhong University of Science and Technology Press.
ISBN of the Co-Publisher's edition: 978-756-80-7743-9

© Huazhong University of Science and Technology Press 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publishers, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publishers nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publishers remain neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Introduction

The course of “Fundamentals of measurement and signal analysis” has been offered to students of School of Mechanical Science and Engineering of Huazhong University of Science and Technology since 1980s. The contents of this course include description and classification of signals, analysis and processing of signals, basic properties of measurement systems, principle of sensors, signal conditioning techniques, engineering applications of information technology. Its core contents are the acquisition, transmission, processing and application of information. In the past two decades, under the guidance of Academician Yang Shuzi, our team has innovated the education mode, focused on the integration of knowledge learning and skill training. This course focuses on teaching the basic theories, techniques and methods related to engineering measurement technology, and also focuses on the practical applications of measurement technology.

The major of Mechanical Engineering is not only a traditional major, but also a cutting-edge interdisciplinary that keeps pace with the time. The manufacturing industry is the principal part of the national economy. Since the beginning of industrial civilization in the mid-eighteenth century, the history of the rise and fall of world powers and China has repeatedly proved that without a strong manufacturing industry, there can be no prosperous of the nation. In the manufacturing industry, measurement technology plays a pivotal role in mechanics, automation and intelligence manufacturing. The textbook lists many high-tech products and major engineering achievements independently developed by China, such as intelligent robot production lines, unmanned warehouses, intelligent CNC machine tools, smart buildings, smart homes, China’s super drilling rig Crust-1, Mars exploration vehicle Tianwen-1 and the manned submersible Fendouzhe. These great achievements would not be accomplished without the measurement technology.

Definition of Measurement Technology

Basic Concepts of Measurement Technology

Measurement is an indispensable part of the development of modern electromechanical equipment and instrument. In engineering measurements, various physical quantities need to be measured to obtain accurate and quantitative results. Measurement is not only needed in various types of engineering tests but also in control, monitoring of machine status and fault diagnosis of machines. Thus, measurement systems are an important part of the machines and production lines. The main content of this book is measurement technology and instruments.

The basic task of measurement is to obtain useful information. Information is always contained in certain physical quantities. The changes of these physical quantities with time are called signals. In terms of physical properties, signals include electrical signals, optical signals, force signals and so on. Among them, electrical signals have obvious advantages in transformation, processing, transmission and application, and thus become the most widely used signal at present. Various non-electrical signals are often converted into electrical signals, and then transmitted, processed and used.

Constitution of Measurement Systems

The Process of Information Acquisition

The complete process of information acquisition is as follows: firstly, detect the relevant information of the measured object, then process the information and finally provide the result to the observer or feed it to other information processing devices and control systems.

It can be seen that measurement technology belongs to the category of information science. Measurement technology, computer technology and communication technology are referred to as the three pillars of information technology.

Measurement technology and measurement systems are everywhere around us. For example, a drone shown in Fig. 1 has a 9-axis motion and orientation sensor, a 3-axis accelerometer, a 3-axis magnetometer, a 3-axis gyroscope, cameras and temperature sensors. These sensors are similar to the sensory organs of humans. They are part of the measurement system that are responsible of converting non-electrical physical quantities into electrical signals.

For a self-driving vehicle shown in Fig. 2, human is no longer a necessary participant. It can start engine, drive and stop automatically. The self-driving vehicle is equipped with cameras, ranging radars and ranging lasers. These sensors are responsible for checking the nearby environment of the road by measuring corresponding



Fig. 1 A drone



Fig. 2 Self-driving vehicle (image courtesy of Prof. Li Weihua of South China University of Technology)

physical quantities. Together with the navigation by GPS sensor, it can fulfill the task of self-driving.



Fig. 3 Mercury and digital thermometers

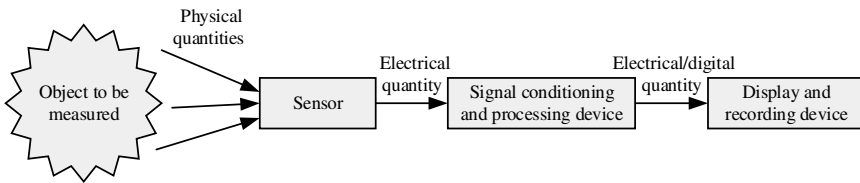


Fig. 4 Schematic diagram of measurement system

Constitution of Measurement System

A simple measurement system has only one module, such as the mercury thermometer shown in the left figure of Fig. 3, which directly converts the temperature value of the measured object into the liquid level. There is no power conversion and signal processing circuit in between. It is relatively simple, but the measurement accuracy is low. At the same time, it is difficult to realize automatic measurement. In order to improve measurement accuracy and realize automatic measurement, in industrial measurement applications, digital thermometer shown in the right figure of Fig. 3 is used, which convert the temperature into electrical quantity, and then process the electrical signal.

Generally speaking, as shown in Fig. 4, the measurement system consists of three parts: sensor, intermediate conversion device, display and recording device. During the measurement, the sensor detects the physical quantities (such as pressure, acceleration and temperature) that reflect the characteristics of the measured object, converts them into electrical signals and then transmits them to the intermediate conversion device; the intermediate conversion device uses analog circuits to process the electrical signal or uses a software to analyze the digital signal after A/D conversion; finally, the display and recording device will display the results to the users or feed the result into other control devices.

(1) Sensor: is a device that can perceive a specific physical quantity and convert it into electrical quantity according to a certain rule. Sensors are usually composed of sensitive elements and conversion structures. When the physical quantity to be measured is difficult to be converted into electrical signals directly, a conversion structure is used to convert the physical quantity into another intermediate physical quantity. Then, the sensing element converts the intermediate physical quantity into electrical signal. It is not necessary for a sensor to have both components, many sensors have only sensitive elements. And for some other sensors, the two components are integrated as one.

(2) Signal conditioning and processing device: converts the sensor output signal into a standardized signal that is convenient for transmission and processing. Because the output signal of sensor is generally weak and noisy, it is not suitable for processing, transmission and recording. Thus, generally, it needs to be modulated, amplified, demodulated and filtered, etc.

Signal conditioning is to convert the signal from the sensor into a form that is more suitable for further transmission and processing, such as amplification, convert the impedance change into a voltage change, or convert the impedance change into a frequency change. Signal processing is to acquire the signal from the conditioning module, and perform various types of calculations, filtering, analysis and then output the results to the display and recording device.

(3) Display and record device: is to display or record the processing results for further analysis by the user. If the measurement system is one part of a control system, the processing results can be further fed into the actuator of the control system.

In measurement systems, there are three words that always make us confusing: sensor, actuator and transducer. The definition of sensor has already been introduced above. An actuator is a device that is responsible for moving and controlling a mechanical structure or a device that converts electrical signal to other types of physical quantity such as force. A transducer is a device that converts energy from one form to another. In some measurement systems, actuators are used to excite signals and sensors are used to receive the signals, while in some other systems, the signals acquired by sensors are fed into actuators for further tasks of control. Certain types of sensors and actuators that have energy conversion can be regarded as transducers such as a piezoelectric disk.

It should be noted that in all the steps of measurement, a basic principle that must be followed is that the output and input of each step should maintain a one-to-one correspondence and the distortion in output should be reduced or eliminated as much as possible.

The composition of measurement system is related to the task, and not necessarily to contain all the modules in Fig. 4. According to the complexity of the measurement task, each module in the measurement system can be composed of multiple sub-modules.

Example 1: Line Tracking Robot Car

Line tracking is one of the control methods for the moving of robots. The intelligent line tracking robot car belongs to the category of robots, as shown in Fig. 5. It

Fig. 5 Intelligent line tracking robot car



integrates mechanics, electronics and control technology, and has broad applications in warehouse intelligent management, high-voltage line inspection, deicing and other fields. The robot car emits infrared light through the excitation circuit, the light is reflected by the ground (white background, black line) and then is received by the sensors. Since the reflection from the white background and the black line are different, the output voltage of the sensor is also different. Thus, the position of the robot car can be determined from the sensors signals, and this information can be used to control the motor to adjust the direction of moving.

Example 2: Noise Measurement

Strong noise will cause negative effects to our physiology and psychology. Noise in the environment mainly causes hearing loss and interferes with talking, thinking, rest and sleep. A noise sensor or microphone can be used to effectively measure the noise. The measured noise intensity can be displayed in real-time on a screen, as shown in Fig. 6.

Engineering Applications of Measurement Technology

Measurement technology has been widely used in various aspects of industry such as agriculture, manufacturing, scientific research, domestic and foreign trade, national defense and transportation. It plays an important role and becomes an essential part in the development of national economy. Therefore, the use of advanced measurement technology has become one of the important signs of technological modernization. In the field of engineering, measurement technology is indispensable for research, product development, production supervision, quality control and performance testing. Even in daily life appliances, such as automobiles, household appliances, measurement technology is an inseparable part. At present, smart manufacturing, smart buildings, smart homes, smartphones, smart medical care, etc., continue to develop. Intelligence has become a major trend in manufacturing and daily life, and the prerequisite of intelligence is sensors and measurement technology. The following are several typical applications of measurement technology.



Fig. 6 Noise measurement

Industrial Automation

In industrial equipment, the measuring device plays the role of perception. For example, the vision sensor, tactile sensor, displacement sensor, force sensor, torque sensor of the robot on the production line, as well as the tracking and positioning sensor of the automatic feeding trolley, as shown in Fig. 7.

Figure 8 shows the measurement of parameters of the production line, including motor current, cutting force, vibration/noise, etc. This information can be used to monitor the cutting process of CNC (computer numerical control) machine tools in real time.

Example 3 Intelligent Machine Tool

Intelligent machine tool is an advanced form of the CNC machine tools. Since traditional CNC machine tools only use G codes and M commands to control the motion of the tool and workpiece, the actual machining status of the machine tool, such as cutting force, inertial force, friction force, vibration, thermal deformation and environmental changes, is not well perceived and used as feedback. Thus, the actual trajectory of the tool deviates from the theoretical path, reducing machining accuracy, surface quality and production efficiency. Therefore, CNC machine tool manufacturers are promoting the CNC machine tools to intelligent machine tools by applying sensor technology, network technology and adding intelligent functions. The in-depth integration of the new generation of artificial intelligence technology and CNC



(a) Intelligent robot production line of Great Wall Motor (image source: find800.cn)



(b) Jingdong unmanned warehouse sorting robot (image source: sohu.com)

Fig. 7 Automated industrial equipment



Fig. 8 State monitoring of CNC cutting process

Fig. 9 Intelligent precision machining center (image courtesy of Huazhong CNC Co., Ltd.)



machine tools will bring new changes to the machine tool industry. The S5H intelligent precision machining center, as shown in Fig. 9, has intelligent function modules such as quality inspection of machine tool assembly, thermal error compensation and contour error compensation. The positioning accuracy can reach $1\text{ }\mu\text{m}$, and the slow feed of the machine tool can reach $1\text{ }\mu\text{m/min}$.

State Monitoring of Industrial Equipment

In many industries such as electric power, metallurgy, petrochemical and chemical industries, the working status of certain key equipment such as steam turbines, gas turbines, water turbines, generators, motors, compressors, fans, pumps, gearboxes affects the running of entire production line. The operating status and change trend of these key equipment can be timely and accurately known by implementing 24-h real-time monitoring which allows the transformation from post-maintenance or regular maintenance to predictive maintenance. Practical experiences show that the vibrations of some important measuring points of the unit can reflect the operating status of the unit. Since most of the unit failures have a gradual development process from quantitative change to qualitative change, by monitoring the vibration of the key parts, the equipment failure can be timely and effectively predicted. Integrate other monitoring information (such as temperature, pressure and flow) to analyze the specific location of the fault, the loss caused by equipment failure can be minimized.

Rotor is an important part of the mechanical system, and its imbalance will cause vibration or even damage to the equipment. Especially for the high-speed rotating rotor, the accident caused by it is more serious. Failures caused by rotor imbalance account for more than 60% of all mechanical failures. With the development of precision CNC machining technology, the dynamic balance problem of high-speed rotor that seriously affects its machining accuracy has attracted particular attention. The vibration of rotating bodies may be caused by the inhomogeneity of material,



Fig. 10 High-speed dynamic balance of G50 gas turbine rotor (image courtesy of Dongfang Turbine Co., Ltd.)

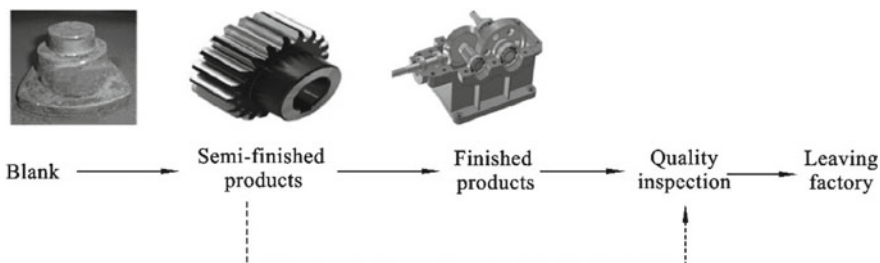


Fig. 11 Quality inspection of gear

defects in blank, errors in process of assembly, asymmetric geometric shapes that exist in the design, as well as various factors such as impact, corrosion and abrasion. The vibration will generate noise, accelerate bearing wear and shorten mechanical life. Therefore, the rotor must be balanced in time to make it reach a desirable accuracy of balance.

Product Quality Inspection

Products such as automobile, machine tool, motor, engine and other parts must be inspected during production and before leaving the factory. In the quality inspection, certain methods are used to measure the product, and measured results are compared with that of standard products to determine if the product is qualified or not. The process of production and inspection of gear is shown in Fig. 11.

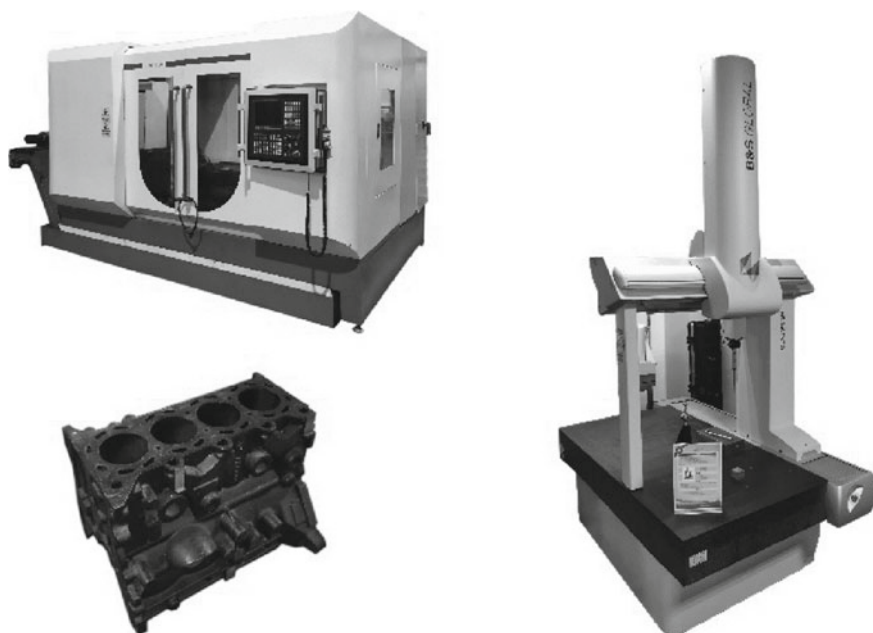


Fig. 12 Measurement of machined parts

Example 4: Measurement of Mechanical Parts

In the measurement of the machining accuracy of mechanical parts, instruments are used to measure their geometrical parameters (such as length and diameter), position parameters (such as coaxiality) and surface quality (such as roughness). Figure 12 shows the measurement of typical machined parts with a three-coordinate measuring machine.

Smart Buildings, Smart Homes and Smart Offices

The smart building is to optimize the structure, system, service and management of the building according to the needs of users through the introduction of sensing and measurement technology. It provides users with an efficient, comfortable and convenient building environment. Smart building is a product of modern science and technology, and its technical foundation is mainly composed of building technology, sensor technology, computer technology, communication technology and control technology. Figure 13 shows a schematic diagram of a typical smart building.



Fig. 13 Smart buildings (image courtesy of Gaozhi Technology Co., Ltd.)

Sensing and measurement technology is also widely used in smart home and smart office to improve the comfort at home and efficiency of work at office. Figure 14 is a schematic diagram of a typical smart home

Measurement Technology in Smartphones

Smartphones have become an important tool in our live due to its various functions. The improvement of cell phone functions is attributed to the sensing and measurement technology. Generally, a smartphone contains a variety of sensors, such as image sensors, microphones, temperature sensors, humidity sensors, gyroscopes, acceleration sensors, pressure sensors, magnetic field sensors, which support useful functions such as photography, navigation and distance measurement.

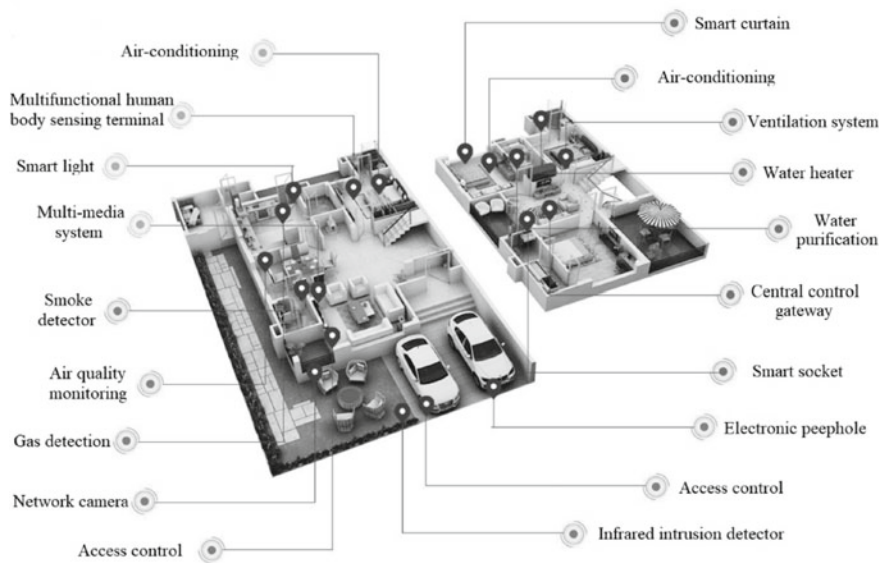


Fig. 14 Smart home (image courtesy of Gaozhi Technology Co., Ltd.)

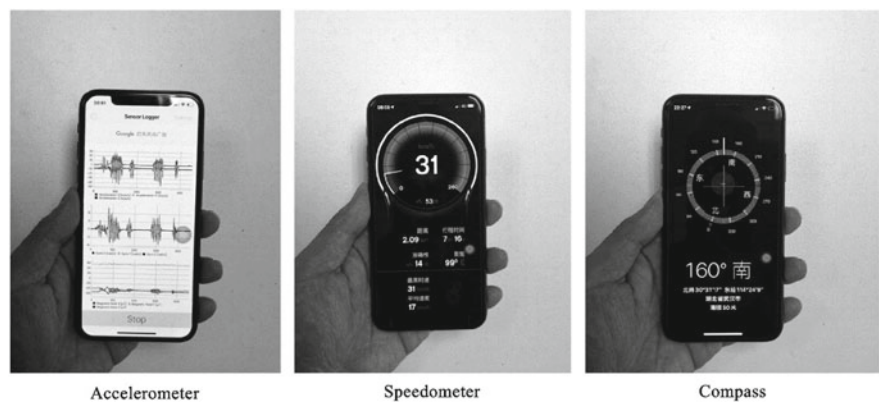


Fig. 15 Sensors in smartphones

Other Applications

Sensing and measurement technology is also widely used in aerospace, smart agriculture, ocean transportation and medical testing, as shown in Fig. 16.

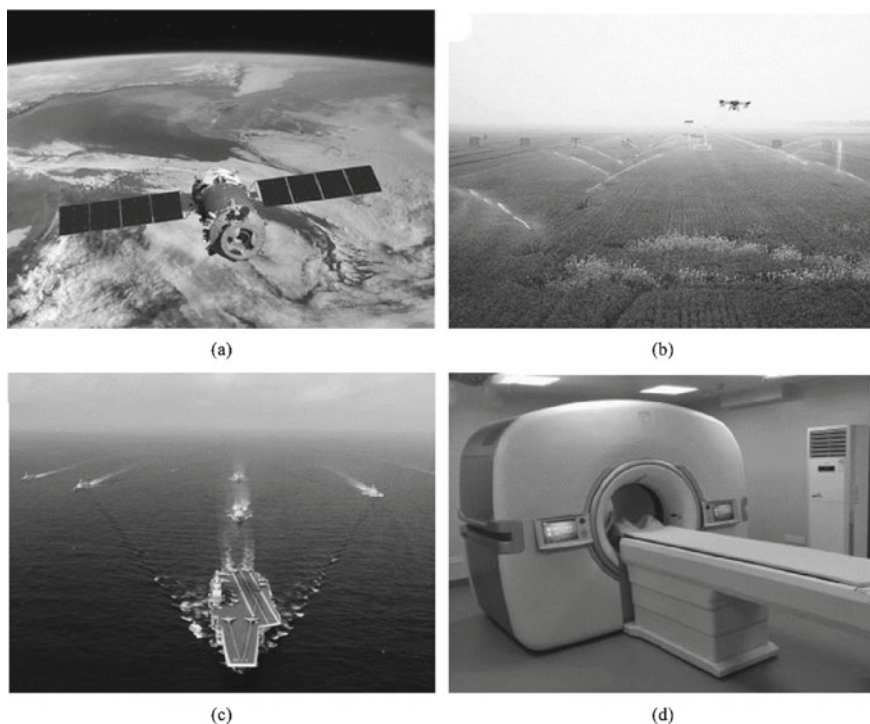


Fig. 16 Applications in the fields of aerospace, agriculture, transportation and medical testing (image source: chinanews.com)

Example 5: Deep Earth Exploration Application.

The deep earth detection instrument is like an eye which can see deep underground minerals and hidden targets on the seabed. It is of great value to homeland security. Professor Huang Danian of Jilin University led a team with more than 400 researchers to develop China's first 10,000-meter drill "Crust-1". It is equipped with a self-developed integrated geophysical data analysis system.

Example 6: Tianwen-1 Mars Probe

"Tianwen-1" is China's first Mars probe developed by China Aerospace Science and Technology Corporation, as shown in Fig. 18. It consists of an orbiter, a lander and a patrol device. It was launched into space by the fourth Long March-5 rocket, breaking through key technologies such as exceeding second cosmic speed launch, interplanetary flight, soft landing on extraterrestrial planets, and completed the missions of orbiting, landing and patrolling at one time. It is the first time in the world's aerospace history to simultaneously achieve these goals.

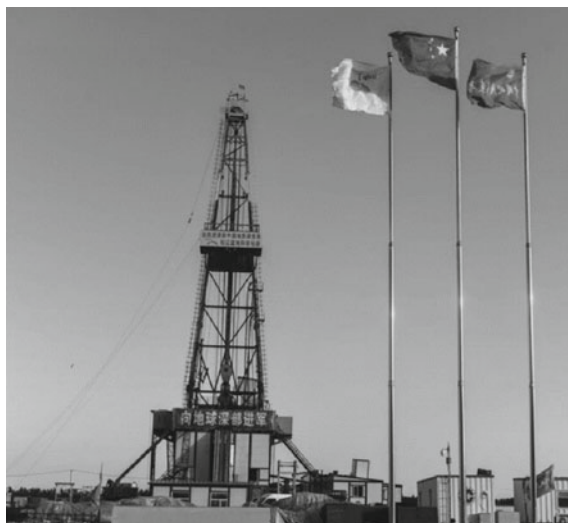


Fig. 17 China's super drill Crust-1 (image source: xinhuanet.com)

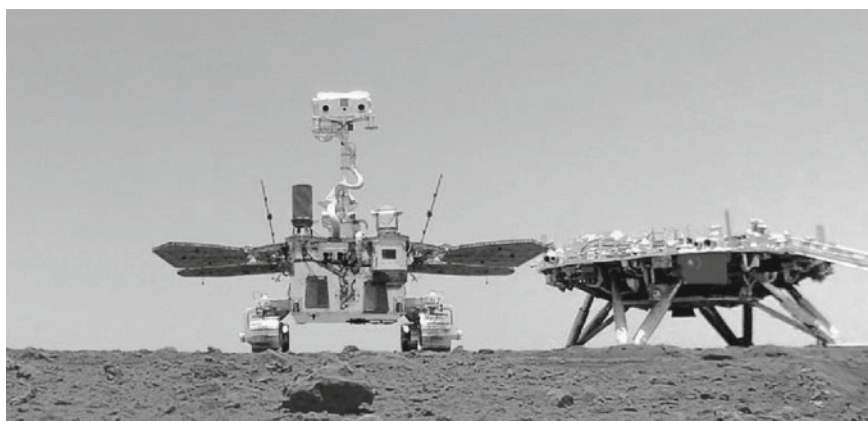


Fig. 18 Tianwen-1 Mars probe (image source: chinanews.com)

Example 7: Fendouzhe Submersible

“Fendouzhe” is another deep sea manned submersible developed by China after the “Jiaolong” and “Deep Sea Warrior”. A scientific research team composed of nearly a thousand researchers from nearly 100 scientific research institutes, universities and enterprises has completed the research. It combines the advantages of the previous two generations of manned submersible of Jiaolong and Deep Sea Warrior. In addition to a safe, stable and powerful energy system, it has a more advanced control system and positioning system, as well as a better pressure-resistant cabin. Starting from October 10, 2020, Fendouzhe went to the Mariana Trench to conduct a 10,000-meter



Fig. 19 Deep sea manned submersible Fendouzhe (image source: chinanews.com)

sea trial and successfully completed 13 dives, of which 8 times exceeded 10,000 meters. At 8:12 on November 10, 2020, Fendouzhe created a new record of 10,909 meters for manned deep exploration in China. It stayed there for 6 h to conduct a series of deep-sea scientific researches and brought back minerals, deep-sea organisms and deep-sea water samples.

Developing Trends of Measurement Technology

At present, the trend of technological development is to be more intelligent. Sensing and measurement are the basis of intelligence. Sensors are developed towards the direction of miniaturization, integration and intelligence. The invention of new sensitive materials and the discovery of new detection mechanisms will strongly promote the sensor technology. Miniaturized sensors have great advantages in special applications, while multi-functional integrated sensors combined with microcomputer chips are intelligent.

Miniaturization, Integration and Intelligence of Sensors

Generally, sensors manufactured by traditional methods are large in size and expensive. With the continuous emergence of new applications (such as the miniaturization

of weapons), the requirements for miniaturization are also put forward for the sensors. The micro-nano machining technology developed from IC (integrated circuit) manufacturing technology is gradually being applied to sensor manufacturing. The size of sensitive element can reach the magnitude of micrometers, sub-micrometers or even nanometers. The miniaturized sensors can be mass-produced now, thus they can be manufactured with cheap price and better performance. The needs in application and the development of manufacturing technology will gradually make the sensors miniaturized.

Besides, with the needs of multi-functional measurement, sensors have also developed from measuring a single physical quantity to the direction of multi-quantity measurement. At present, measurement technology is developing in the direction towards multi-function, integration and intelligence. The dynamic measurement of fast-changing parameters is an important part of the automated control system, and its main pillars are microelectronics and computer technology. The sensor is combined with the microcomputer chip to develop a smart sensor. It has the functions of automatic range and gain selection, automatic real-time calibration, non-linear correction, drift error compensation and even complex calculation operations. It can complete automatic fault diagnosis, overload protection, communication and control.

High Precision, High Speed, Large Range

Accuracy is an eternal topic of measurement technology. With the development of science and technology, the requirements for measurement accuracy are getting higher and higher. In the field of size measurement, the requirements for measurement accuracy have reached nanometer and sub-nanometer; in time measurement, the resolution is required to reach femtoseconds, and the relative accuracy is 10^{-14} ; in terms of electric charge measurement, it is required to measure the charge of a single electron; in the aerospace field, the required measurement accuracy for velocity and acceleration have reached a relative error of 0.05%. In terms of speed measurement, the running speed of machine tools, turbines, vehicles, etc., are all become larger, the rotation speed of the turbine rotor has reached hundreds of thousands of times per minute. To complete the accurate measurement of the air gap between the turbine rotor and the stator, the sampling time must be femtosecond and the sampling rate must be at least terahertz. High-tech fields such as national defense and aerospace have even higher requirements for speed measurement. The aircraft must continuously correct its trajectory, attitude and acceleration during operation, which requires rapid response of the measurement instrument. When launching a rocket, the measurement of speed should be accurate and rapid to avoid serious disasters. In the research of explosion and nuclear reaction, the measurement instrument is often required to be able to respond in microseconds.

In the process of scientific and technological progress, new fields and new things are constantly emerging. Therefore, measurement technology is in need in more and more fields. Some measurement conditions are getting worse and worse, such as high

temperature, high speed, high humidity, high density of dust, large vibration, high pressure, high voltage, deep water, strong field, explosive and so on. There are more and more types of parameters that need to be measured. Some require the realization of networked measurement in order to achieve simultaneous measurement in cross-regional situations. Some require simultaneous measurement of multiple parameters, and the synchronization requirements reach the microsecond level. All of these require more powerful measurement methods. Conventional measurement has been relatively mature, but with the continuous deepening of research and the continuous expansion of the field, measurement tasks under some extreme conditions have also continued to emerge. In size measurement, it is required to measure dimensions from nuclei to space; in voltage measurement, it is required to measure dimensions from nanovolts to millions of volts; in resistance measurement, it is required to measure dimensions from superconductivity to $10^{14} \Omega$; in acceleration measurement, it is required to measure dimensions from $10^{-4} g$ to $10^4 g$ (g is the gravitational acceleration); in temperature measurement, it is required to measure dimensions from close to absolute zero to $10^{18} K$. Measurement technology is developing in the direction of solving these extreme measurement problems.

Network-Based Measurement and Automatic Measurement

With the rapid development of computer technology, communication technology and network technology, networked measurement technology based on computers and workstations has become a new development trend. Bus is a part of the network, which can be connected to the Internet and local area network, and has an open and unified communication protocol. So it is responsible for the tasks of measurement and control in manufacturing processes. The networked measurement system generally consists of a measurement module, a signal transmission module, and a signal analysis and processing module. The outstanding feature of networked measurement technology is that it can realize resource sharing, multi-system, multi-task, multi-expert collaborative measurement and diagnosis. Technicians can obtain required information anytime and anywhere.

The preparation time for a large-scale measurement is long, and there are many parameters to be tested. If data are processed manually, not only the accuracy is low, but the processing cycle is also too long. The development of modern measurement technology makes it possible to adopt an automated measurement system with a computer as the core. The system can realize automatic calibration, automatic correction, fault diagnosis, signal modulation, multi-channel acquisition automatic analysis and processing, and can print out the measurement results. At the same time, there is also a measurement technology that directly embeds micro-measurement system into the measured object. The system measures the changes of various parameters of the measured object during its operations and stores the data. Then the stored data can be read through computer interface.

Intelligent Processing of Measurement Signal

Before the 1950s, the measurement signals were mainly analyzed by analog analyzing methods. After the 1950s, computers had practical applications in signal analysis. At that time, researchers had argued about the advantages and disadvantages of analog and digital analysis methods. The focus of the debate was speed, accuracy and cost of these processing methods. In the 1960s, the development of artificial satellites, aerospace explorer, communication technology and radar put forward higher requirements for the speed and resolution of signal analysis. In 1965, J. W. Cooley and J. W. Tukey proposed the Fast Fourier Transform algorithm, so that the calculation time of Discrete Fourier Transform has been reduced from N^2 to $N \cdot \log_2 N$. This method promotes the digital signal processing and makes it more widely used. After the 1970s, the development of large-scale integrated circuits and the application of microcomputers further promote the technology of signal analysis. Many algorithms appeared afterwards. In 1968, C. M. Rader proposed the Number theoretic transforms FFT algorithm (NFFT); in 1976, S. Winograd proposed a Winograd Fourier Transform Algorithm (WFTA), which reduced the number of multiplication operations to 1/3 of the FFT algorithm; in 1977, H. J. Nussbaumer proposed a Polynomial Fourier Transform Algorithm (PFTA), which combines FFT and WFTA methods to increase the speed of the FFT algorithm by about 3 times when the number of sampling points is large.

With the advent of the era of big data and artificial intelligence, new data processing methods represented by machine learning have been gradually proposed. Machine learning is a multi-disciplinary interdisciplinary, involving probability theory, statistics, approximation theory, convex analysis, algorithm complexity theory and other disciplines. It specializes in the simulation of human learning behaviors with computers. With the continuous increase in the demand for data analysis in various industries, the efficient acquisition of knowledge through machine learning has gradually become the main driving force for the development of today's machine learning technology. Machine learning becomes a support and service technology for many fields. Performing in-depth analysis of complex data based on machine learning has become the main direction of the research of machine learning. Machine learning has become an important mean of intelligent data analysis. In addition, in the era of big data, as the speed of data generation continues to accelerate, the volume of data has increased unprecedentedly. New types of data that need to be analyzed are also emerging, such as text understanding, text sentiment analysis, image comprehension, analysis of graphics and network data. This makes intelligent computing technologies such as machine learning and data mining play an extremely important role in the application of big data analysis and processing. Common machine learning algorithms include decision trees, Naive Bayes, support vector machines, random forests, artificial neural networks and so on.

Deep learning is a new research filed of machine learning. It is a representation learning method. The learning is performed in an artificial neural network with several layers as shown in Fig. 20. In supervised learning, the leaning is accomplished by

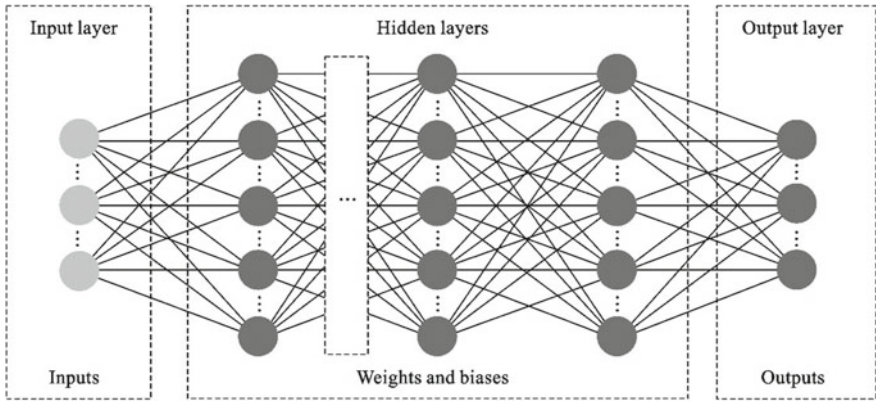


Fig. 20 An artificial neural network for deep learning

feeding a large number of input-output pairs to adapt the weights in each neuron. Deep learning is to extract the internal law and expression of the sample data. The information obtained in the learning process is of great help to the interpretation of data such as text, image and sound. The goal is to allow machines to have the ability to analyze and learn like humans, and to recognize data such as text, images and sounds. Deep learning is a very complex machine learning algorithm, and its effects in speech and image recognition far surpass previous related technologies. Deep learning has completed many achievements in searching, data mining, translation, natural language processing, multimedia learning, personalized recommendation and other related fields.

Main Manufacturers of Sensor and Measurement Instrument

(1) Main manufacturers and their products

BK PRECISION mainly provides vibration and noise measurement sensors and instruments. National Instruments mainly provides virtual instrument software such as LabVIEW, BridgeVIEW and measurement hardware such DAQ. Agilent (HEWLETT PACKARD) provides a wide range of measurement instrument such as oscilloscope and function generator. KEITHLEY provides a wide range of measurement instrument, data acquisition instruments and semiconductor devices. Tektronix provides a wide range of instruments, especially oscilloscopes, function generators and VXI bus instruments. Telulex provides arbitrary waveform generators (AM, FM, PM, SSB, BPSK DC-20 MHz). ADVANTEST provides IC testers, digital voltmeters. Iotech provides data acquisition and PC-based instruments. PICO

provides virtual instruments and PC-based digital oscilloscopes. Apogee Instruments provides CCD cameras and other CCD device-based instruments, as well as some graphics processing software. Amplifier Research mainly provides broadband RF power amplifiers. ANCOT provides fiber channel and small computer system interface; ANRITSU provides radio communication analysis instruments and spectrum analyzers. Applied Microsystem provides software and hardware for embedded systems, supporting chips provided by Intel, AMD and Motorola. Berkeley Nucleonics provides pulse generators. IWATSU provides oscilloscopes and digital storage oscilloscopes. LeCroy provides digital storage oscilloscopes, function generators. NF Instruments provides signal generators, distortion meters, acoustic emission detectors and various amplifiers. Nicolet Instrument Technologies, Panasonic, Yokogawa mainly provide test instruments. Stanford Research Systems provides FFT analyzers.

(2) Main electronic resources and magazines

IEEE/IEE Electronic Library, Wiley Online Library, Energy Information Administration of America, Test and Measurement World Online, Sensors and Actuators B: Chemical, Sensors and Actuators A: Physical, IEEE/ASME Transactions on Mechatronics, Measurement, Frontiers of Mechanical Engineering

Exercise

1. Discuss the basic structure of a measurement system and the basic functions of each module of the system.
2. What is the basic idea of the non-electrical quantity measurement?
3. List some measurement applications in engineering to illustrate the role and status of informatization and digitization of measurement instrument.
4. Summarize the process and method of measurement according to the measurement process of a certain physical quantity that you are familiar with.
5. Give an example to illustrate the main difference between direct measurement and indirect measurement.
6. According to your understanding, talk about the meaning of measurement technology.

Contents

1	Waveform Analysis of Signals	1
1.1	Classification of Signals	1
1.1.1	Deterministic and Non-Deterministic Signal	2
1.1.2	Energy Signal and Power Signal	4
1.1.3	Other Classifications of Signals	6
1.2	Sampling Theorem	10
1.2.1	A/D Conversion	10
1.2.2	Sampling Error	11
1.2.3	Nyquist–Shannon Sampling Theorem	12
1.2.4	Frequency Aliasing in Sampling	12
1.2.5	Anti-Aliasing Filtering Before A/D Conversion	15
1.2.6	Quantization Error	16
1.2.7	Technical Index of A/D Converter	17
1.2.8	D/A Conversion	19
1.3	Standard Functions in Signal Analysis	20
1.3.1	Unit Impulse Function (Δ Function)	20
1.3.2	Unit Step Function	22
1.3.3	Unit Ramp Function	23
1.3.4	Complex Exponential Function	24
1.3.5	Sine Function	26
1.3.6	Sinc Function	26
1.3.7	White Noise	27
1.4	Generation of Standard Functions	28
1.5	Waveform Analysis of Signals	32
2	Frequency Domain Analysis	39
2.1	Concept of Frequency Domain Analysis	39
2.1.1	Advantages of Frequency Domain Analysis	41
2.1.2	Objectives of Frequency Domain Analysis	43
2.2	Fourier Series Representation of Periodic Signals	44
2.2.1	Orthogonal Decomposition of Vectors	44

2.2.2	Orthogonal Function	45
2.2.3	Orthogonality of Trigonometric Functions	46
2.2.4	Fourier Series in Trigonometric Function Form	48
2.2.5	Spectrum of Periodic Signals	50
2.2.6	Synthesis of Periodic Signals	53
2.2.7	Walsh Orthogonal Function Set	55
2.2.8	Fourier Series in Complex Form	56
2.2.9	Gibbs Phenomenon	59
2.2.10	Parseval's Theorem	60
2.3	Fourier Transform of Aperiodic Signals	61
2.3.1	Fourier Integral	61
2.3.2	Spectrum of Aperiodic Signal	63
2.4	Fourier Transform of Typical Signals	64
2.5	Properties of Fourier Transform	72
2.6	Fourier Transform for Discrete-Time Signals	75
2.6.1	Sampling in Time and Frequency Domain	76
2.6.2	Discrete Fourier Series (DFS)	81
2.6.3	Discrete Fourier Transform (DFT)	82
2.6.4	Fast Fourier Transform (FFT)	85
2.6.5	Applications of FFT Algorithm	87
2.7	Error Analysis in FFT	96
2.7.1	Signal Truncation	96
2.7.2	Effect of Spectral Leakage	97
2.7.3	Fence Effect	98
2.7.4	Window Functions	99
2.8	Applications of Frequency Domain Analysis	103
3	Amplitude Domain Analysis	109
3.1	Probability Density Function and Histogram	109
3.2	Probability Distribution Function	115
3.2.1	Cumulative Distribution of Discrete Random Variables	116
3.2.2	Probability Distribution of Continuous Random Variables	116
3.3	Engineering Applications of Amplitude Domain Analysis	120
3.3.1	Machine Fault Diagnosis	120
3.3.2	Histogram Analysis of Photo Quality	122
4	Correlation Analysis of Signals	125
4.1	Concept of Correlation Analysis	125
4.1.1	Correlation of Variables	125
4.1.2	Correlation of Signals	126
4.1.3	Cross-Correlation Function	127
4.1.4	Auto-Correlation Function	127
4.1.5	Convolution	128
4.1.6	Convolution Theorem	128

4.2	Properties of the Correlation Function	129
4.2.1	Properties of Auto-Correlation Function	129
4.2.2	Properties of Cross-Correlation Function	131
4.2.3	Convolution, Correlation and Fourier Transform	132
4.3	Calculation of Correlation	134
4.3.1	Numerical Calculation	134
4.3.2	FFT Method	135
4.4	Engineering Applications of Correlation Analysis	138
5	Time–Frequency Domain Analysis	143
5.1	Motivations of Time–Frequency Domain Analysis	143
5.1.1	Non-stationary Signals	143
5.1.2	Drawbacks of Global Analysis of Non-stationary Signals	143
5.1.3	Time–Frequency Domain Analysis	144
5.2	Short-Time Fourier Transform	145
5.2.1	Basic Principle of Short-Time Fourier Transform	145
5.2.2	Time–Frequency Resolution of Short-Time Fourier Transform	149
5.2.3	Time–Frequency Domain Decomposition and Synthesis	150
5.2.4	Discrete Short-Time Fourier Transform	150
5.2.5	Methods to Improve Estimation of Spectrogram	152
5.3	Wavelet Transform	153
5.3.1	Wavelet Transform and Short-Time Fourier Transform	153
5.3.2	Continuous Wavelet Transform	155
5.3.3	Applications of Wavelet Analysis	159
6	Digital Filters	161
6.1	Concept of Filtering	161
6.1.1	Ideal Filters	164
6.1.2	Practical Filter	167
6.1.3	Digital Filters	169
6.2	Filtering in Frequency Domain	172
6.3	Time Domain Filtering and Z-transform	177
6.3.1	Time Domain Filtering	177
6.3.2	Z-transform	180
6.3.3	Bilateral Z-transform	182
6.4	Finite Impulse Response (FIR) Filter	182
6.5	Infinite Impulse Response (IIR) Filter	184
6.6	Exercise	186

7	Principles of Sensors	189
7.1	Overview of Sensor Technology	189
7.1.1	Definition of Sensor	189
7.1.2	Composition of Sensors	190
7.1.3	Classifications of Sensors	191
7.2	Resistive Sensors	192
7.2.1	Potentiometer	193
7.2.2	Resistive Strain Gauge	197
7.2.3	Other Resistive Sensors	204
7.3	Inductive Sensors	211
7.3.1	Self-inductive Sensor	212
7.3.2	Eddy Current Sensor	219
7.3.3	Mutual-Inductive Sensor	220
7.4	Capacitive Sensors	223
7.5	Magnetoelectric Transducer	227
7.5.1	Moving Coil Type	228
7.5.2	Variable Reluctance Type	229
7.6	Piezoelectric Transducer	231
7.6.1	Piezoelectric Element	231
7.6.2	Ultrasonic Transducer	235
7.6.3	QCM Humidity Sensor	239
7.7	Hall Sensor	240
7.8	Photovoltaic Transducer	243
7.9	Image Sensors	245
7.10	Thermocouple	250
7.10.1	Thermoelectric Effects	250
7.10.2	Thermoelectric Laws	252
7.11	Fiber-Optic Sensor	256
7.11.1	Frequency/Wavelength Type	257
7.11.2	Intensity Type	258
7.12	Grating Sensor	260
7.13	Biosensor	262
7.13.1	Enzymatic Sensor	262
7.13.2	Microorganism Sensor	263
7.13.3	Immunosensor	265
7.14	Selection of Sensors	266
8	Signal Conditioning Techniques	271
8.1	Overview of Signal Conditioning	271
8.2	Analog Amplifiers and Operators	272
8.2.1	Operational Amplifier	272
8.2.2	Typical Amplification and Operation Circuits	273
8.3	Bridge Circuits	284
8.3.1	DC Bridge	285
8.3.2	AC Bridge	287

- 8.4 Analog Filters 292
 - 8.4.1 Low-Pass Filter 292
 - 8.4.2 High-Pass Filter 296
 - 8.4.3 Band-Pass Filter 298
 - 8.4.4 Band-Stop Filter 300
 - 8.4.5 Comparison of Digital Filters and Analog Filters 301
- 8.5 Modulation and Demodulation 303
 - 8.5.1 Amplitude Modulation 303
 - 8.5.2 Frequency Modulation and Phase Modulation 312
- 9 Characteristics of Measurement System 317**
 - 9.1 Overview of Measurement System 317
 - 9.2 Static Characteristics of Measurement Systems 317
 - 9.3 Dynamic Characteristics of Measurement Systems 320
 - 9.3.1 Transfer Function and Frequency Response Function 320
 - 9.3.2 Linear Measurement System 322
 - 9.4 Characteristics of Typical Linear Measurement Systems 325
 - 9.4.1 Zero Order System 327
 - 9.4.2 First Order System 327
 - 9.4.3 Second Order System 330
- 10 Computerized Measurement System 335**
 - 10.1 Overview of Computerized Measurement System 335
 - 10.2 Development of Measurement Instrument 338
 - 10.3 Computerized Measurement Instrument 339
 - 10.3.1 Virtual Instrument 339
 - 10.3.2 Network Based Measurement Instrument 343
 - 10.3.3 Internet of Things (IoT) 349

Chapter 1

Waveform Analysis of Signals



Signal is the carrier of information. Generally speaking, signals can be classified as electrical and non-electrical signals according to the physical properties. Non-electrical signals include optical signals, acoustic signals, etc. For example, our ancestor lit branches in the beacon to generate smoke, which is an optical signal conveying the information of enemy's invasion. When we speak, the sound wave arriving at others' ears is an acoustic signal carrying the information that we want to express. Electrical signals are voltage and current signals, they are the signals that are easiest to be transmitted and processed nowadays. The signals discussed in this book mainly refer to the electrical signals. In industrial measurement, we use sensors to obtain electrical signals by converting other physical quantities such as force, temperature and vibration into voltages and currents. The recorded change of electrical quantity with time is called measurement signals. They reflect the change of measured physical quantity and contain useful information of the measured object (Fig. 1.1).

The waveform of a signal refers to the change of its strength with respect to time. Its x -axis is time and y -axis is signal strength (amplitude). Signal waveform can be observed in an oscilloscope.

1.1 Classification of Signals

Measurement signals exhibit different characteristics for different physical quantities to be measured. Before we process measurement signals, their characteristics must be known for us to choose a proper signal processing method. Thus, we must know the classifications of signals. According to different classification rules, signals can be classified into different types. From the perspective of mathematical description, signals can be classified as deterministic signal and non-deterministic signal; from the perspective of amplitude and energy, signals can be classified as energy

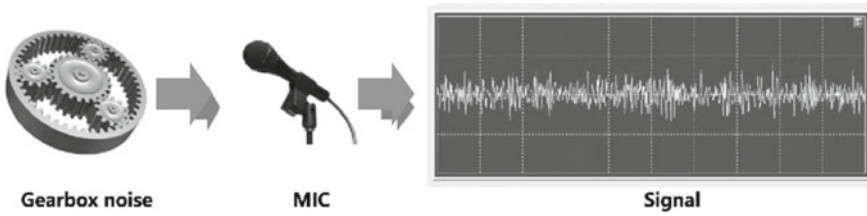


Fig. 1.1 Noise signal of a gear

signal and power signal; from the perspective of analysis domain (abscissa, independent variable), signals can be classified as time-domain finite signals and frequency-domain finite signals; from the perspective of continuity, signals can be classified as continuous-time signal and discrete-time signal; from the perspective of causality, signals can be classified as causal signal and non-causal signal.

1.1.1 Deterministic and Non-Deterministic Signal

1. Deterministic Signal

A deterministic signal is a signal that can be expressed by mathematical equations or certain waveforms, i.e. the signal value at any given time t is deterministic. Many signals appeared in physical processes are deterministic signals, such as the change of voltage of a capacitor during charging and discharging. Deterministic signal can be further classified as periodic signal and aperiodic signals.

(1) Periodic signal

A periodic signal is a signal that repeats its values at regular intervals, which can be mathematically expressed as:

$$x(t) = x(t + nT), \quad n = \pm 1, \pm 2, \dots \quad (1.1)$$

where t is time and T is the repeating period. For example, the vibration caused by an unbalanced rotating body in mechanical systems is often a periodic signal.

Periodic signal can be further classified as simple periodic signal and complex periodic signal. Examples of simple periodic signal and complex periodic signal are given in Fig. 1.2a, b. A simple periodic signal is a single frequency sinusoid, e.g. $x = 0.5\sin(2\pi \cdot 500t)$ and a complex periodic signal is the summation of sinusoids with different frequencies, e.g. $x = 0.5\sin(2\pi \cdot 500t) + 0.5\cos(2\pi \cdot 1000t) + 0.5\cos(2\pi \cdot 1500t)$.

Example 1.1 Generate a simple periodic signal The following MATLAB code can be used to generate a simple periodic signal with frequency of 500 Hz.

```

Fs = 44,100; % sampling frequency
N = 44,100; % number of sampling points
T = 1; % signal duration
x = linspace(0,T,N);
y = 0.5*sin(2*3.14*500*x);
plot(x,y)
xlim([0,0.01])
ylim([-1,1])

```

Example 1.2 Generate a complex periodic signal The following MATLAB code can be used to generate a complex periodic signal with frequencies of 500 Hz, 1000 Hz and 1500 Hz.

```

Fs = 44,100;
N = 4096;
T = 0.01;
x = linspace(0,T,N);
y1 = 0.5*sin(2*3.14*500*x);
y2 = 0.5*cos(2*3.14*1000*x);
y3 = 0.5*cos(2*3.14*1500*x);
y = y1 + y2 + y3;
plot(x,y,'b','linewidth',1)
xlim([0,0.01])
ylim([-1.5,1.5])

```

(2) Aperiodic signal

An aperiodic signal is a signal that does not show repeating characteristics, such as the impact force generated by a hammer, the change of stress in a rope when it is breaking, and the change of water temperature when it is heated. These signals can be expressed by mathematical equations, thus they are deterministic signals. However, they are aperiodic because there is no repeating sequence in the signals.

Aperiodic signals can be further classified into quasi-periodic signal and temporal signal. If a signal is the summation of several period signals and the periods of the periodic signals do not have a least common multiple, then the synthesized signal is no longer periodic. However, the signal still has the characteristic of discrete frequency spectrum, which is a unique feature of periodic signals, thus this kind of signal is called quasi-periodic signal. As shown in Fig. 1.3a, $x(t) = \sin(2\pi \cdot t) + \sin(2\pi \cdot \sqrt{2} \cdot t)$ is a quasi-periodic signal, it is the summation of two periodic signals with period of 1 and $1/\sqrt{2}$. The ratio of the two periods is an irrational number, thus we cannot find such integers n that satisfy Eq. (1.1). Quasi-periodic signals often appear in communication and vibration systems, and are used in mechanical rotor vibration analysis, gear noise analysis and voice analysis. A temporal signal is a signal that lasts in a finite time interval, such as $x(t) = e^{-t} \sin(3t)$ shown in Fig. 1.3b.

Example 1.3 Generate a quasi-periodic signal The following MATLAB code can be used to generate a quasi-periodic signal with frequencies of 500 Hz and $500/\sqrt{2}$ Hz.

```

Fs = 44,100;
N = 4096;
T = 0.01;
x = linspace(0,T,N);
f1 = 500;
f2 = sqrt(2)*500;
y1 = 0.5*sin(2*3.14*f1*x);
y2 = 0.5*cos(2*3.14*f2*x);
y = y1 + y2;
plot(x,y,'b','linewidth',1)
xlim([0,0.01])
ylim([-1.5,1.5])

```

2. Non-Deterministic Signal

If a signal describes a random process, its amplitude and phase are unpredictable. Then it cannot be expressed by mathematical equations. We need to use statistical method to describe it. This type of signal is called non-deterministic signal or random signal. Non-deterministic signals widely exist in our daily life and industries, such as the vibration of a moving car, the noise generated during machining and the trajectory of a falling leaf. Non-deterministic signals can be further divided into stationary random signal and non-stationary random signal. If the mean value, variance, central frequency and other statistic values do not change with time, then it is called stationary random signal, as show in Fig. 1.4a. Otherwise, it is called non-stationary random signal, as show in Fig. 1.4b.

It needs to be pointed out that the actual physical process is often very complicated. There is neither ideal deterministic signal nor ideal non-deterministic signal. They are usually mixed with each other. The classification of signals is shown in Fig. 1.5.

1.1.2 Energy Signal and Power Signal

1. Energy Signal

In engineering measurement, physical quantities to be measured are always converted to electrical signals by sensors. When a voltage signal $x(t)$ is applied to a resistor of $1\ \Omega$, the instantaneous power is:

$$P(t) = \frac{x^2(t)}{R} = x^2(t) \quad (1.2)$$

The integral of instantaneous power over the time interval (t_1, t_2) is the energy consumed by the resistor:

$$E(t) = \int_{t_1}^{t_2} P(t)dt = \int_{t_1}^{t_2} x^2(t)dt \quad (1.3)$$

For signals in a general sense, dimensionless values are used in the calculations of power and energy.

If the signal has finite energy in the interval of $(-\infty, \infty)$, namely it satisfies the following equation:

$$\int_{-\infty}^{\infty} x^2(t)dt < \infty \quad (1.4)$$

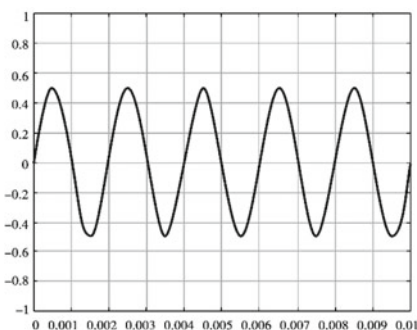
then the signal $x(t)$ is called energy signal. Usually, a temporal signal with finite duration is an energy signal, such as the signal shown in Fig. 1.3b.

2. Power Signal

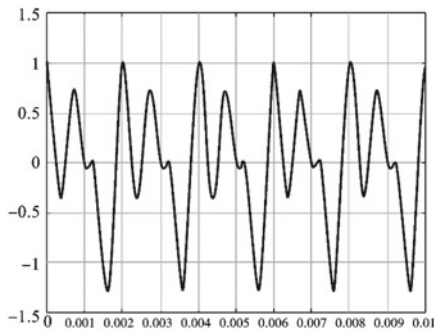
If a signal $x(t)$ has infinite energy in the interval $(-\infty, \infty)$, while it has finite power within a finite interval $(-T/2, T/2)$, then it is called power signal. Power signal should satisfy the following equation:

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t)dt < \infty \quad (1.5)$$

Usually, signals with infinite duration (e.g. periodic signal) are power signals, such as the one shown in Fig. 1.2.



(a) A simple periodic signal.



(b) A complex periodic signal.

Fig. 1.2 Periodic signals

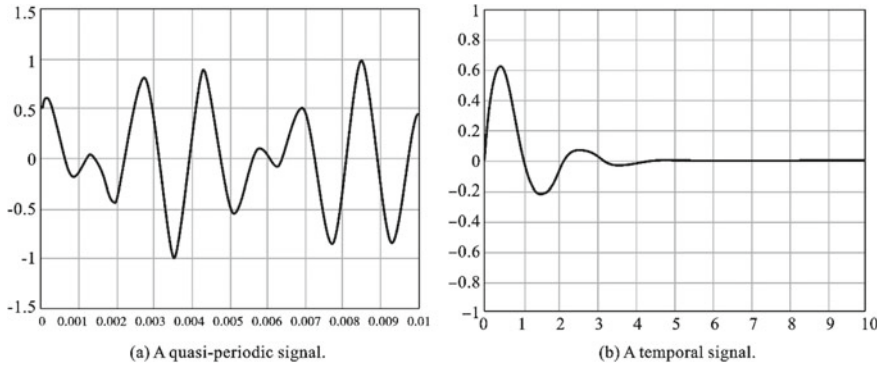


Fig. 1.3 Aperiodic signals

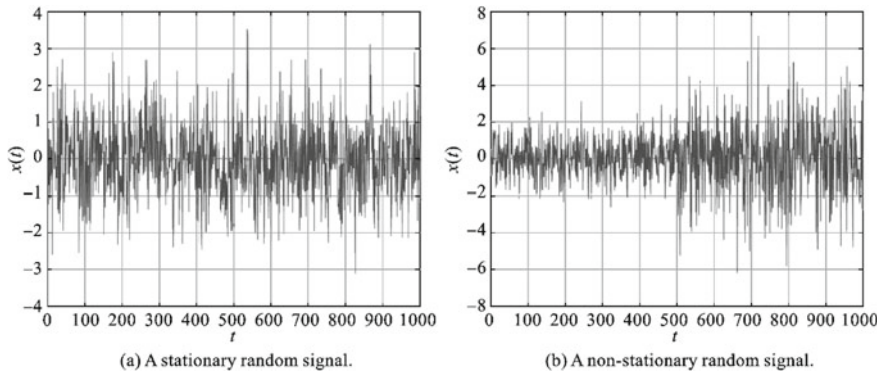


Fig. 1.4 Non-deterministic signals

1.1.3 Other Classifications of Signals

1. Time-Limited and Band-Limited Signals

(1) Time-limited signal

A signal is said to be time-limited if it only has nonzero values for a finite time interval (t_1, t_2) , such as triangular impulse signal, rectangular impulse signal and explosion signal as shown in Fig. 1.6a.

(2) Band-limited signal

A signal is said to be band-limited if it only has nonzero values for a finite frequency interval (f_1, f_2) after Fourier transform, such as sinusoidal signal and signals after band-pass filtering. An example is given in Fig. 1.6b.

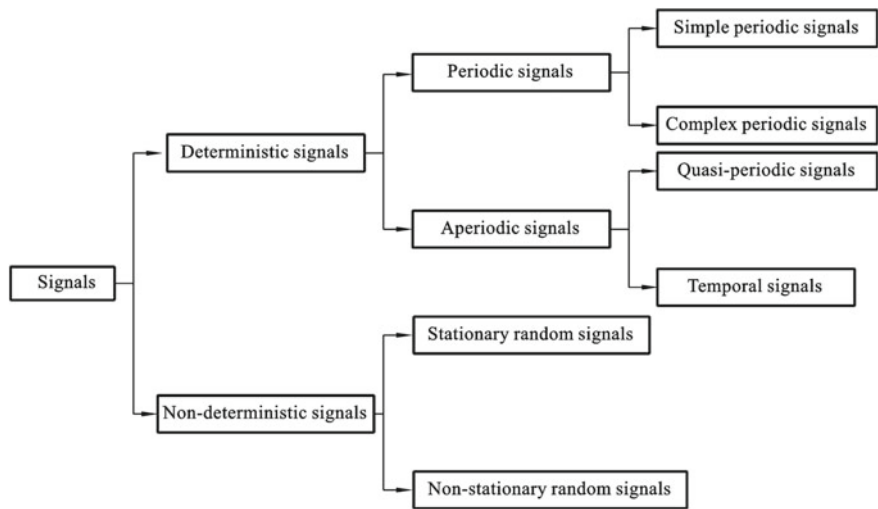


Fig. 1.5 Classification of signals

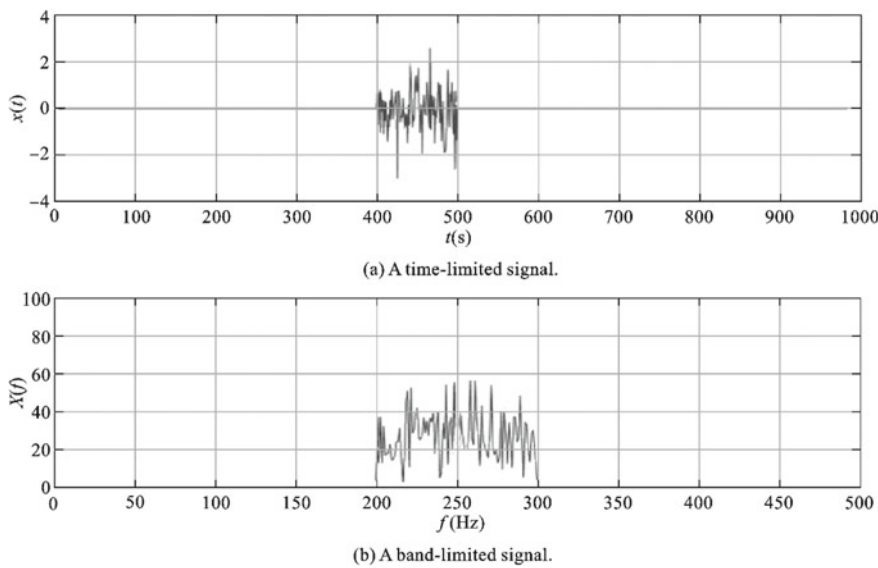


Fig. 1.6 Time-limited and band-limited signals

According to the theory of Fourier transform, a time-limited signal must be infinite in frequency domain, and vice versa. Obviously, a signal cannot be time-limited and band-limited at the same time. Detailed explanations will be given in Chap. 2.

2. Continuous-Time and Discrete-Time Signals

A continuous-time signal is a signal that has definition at any instant of time t within an interval, while a discrete-time signal is a signal that has definition at discrete instants of time n . Examples of continuous-time and discrete-time signals are given in Fig. 1.7.

Considering the continuity the signal amplitude, the signals can be further classified as: continuous-amplitude and continuous-time signal, continuous-amplitude and discrete-time signal, discrete-amplitude and continuous-time signal, discrete-amplitude and discrete-time signal. For example, the analog signal recorded by a microphone is a continuous-amplitude and continuous-time signal; the music stored in computer is a discrete-amplitude and discrete-time signal (Fig. 1.8).

3. Physically Realizable and Physically Unrealizable Signals

Physically realizable signal is also called one-sided signal, it should satisfy the condition of $x(t) = 0$ when $t < 0$. Namely, the signal is completely determined by the side

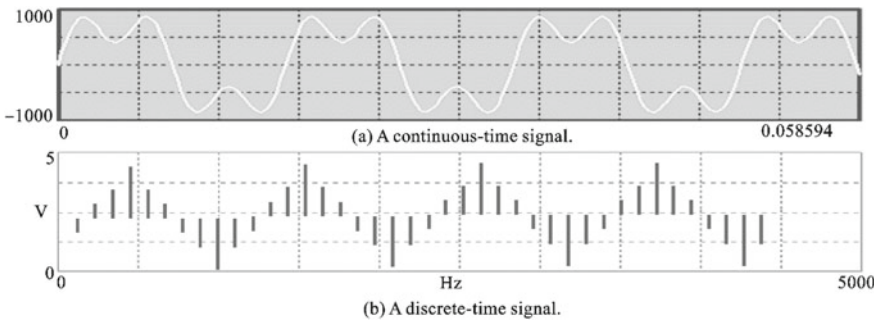


Fig. 1.7 Continuous-time and discrete-time signals

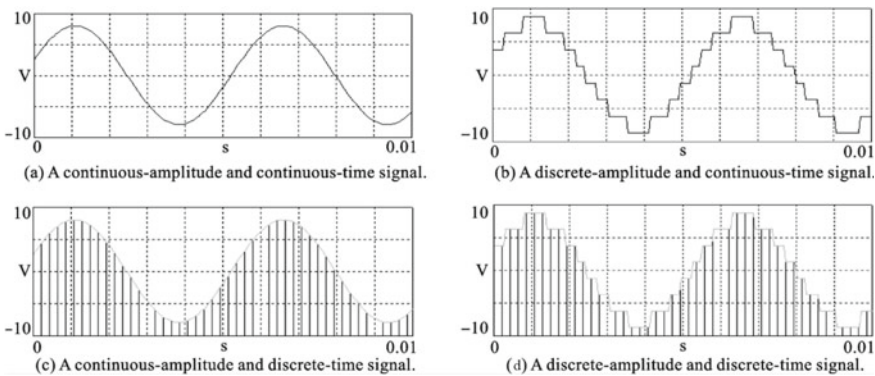


Fig. 1.8 Continuous and discrete signals

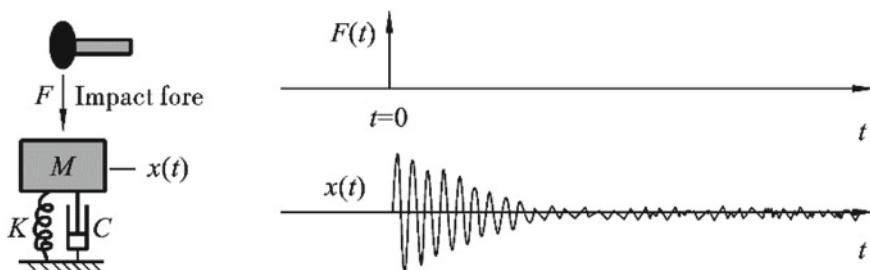


Fig. 1.9 Physically realizable signal

where t is greater than zero, as shown in Fig. 1.9. Otherwise, it is called physically unrealizable signal. Most signals in engineering measurement are physically realizable signals.

Many practical engineering signals are generated when an impulse input is fed into a physical system. For example, in the cutting process, the system composed of machine tools and workpieces can be regarded as a physical system. The sudden change of the cutting tool edge can be treated as the impulse vibration source. Only after the impulse acts on the system, there will be an output in the sensor to indicate the vibration. The so-called physical system has the characteristic that there is no system response until an input is fed into the system. In other words, the output is zero before an input is applied. Thus, for a signal to be realized by a physical system, it must satisfy the condition of $x(t) = 0$ when $t < 0$. Similarly, for discrete-time signals, a sequence that satisfies the condition of $x[n] = 0$ ($n < 0$) is called a physically realizable sequence.

Physically unrealizable signals do not exist in reality. It requires the system to respond before the event occurs ($t < 0$), and this is unrealizable in physics. However, this type of signal exists theoretically. For example, for a rectangular transfer function and a unit impulse, there will be response before the impulse happens, as shown in Fig. 1.10.

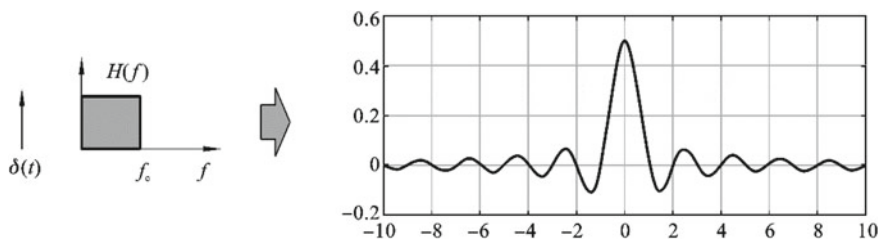


Fig. 1.10 Physically unrealizable signal

1.2 Sampling Theorem

Sampling theorem was proposed by American engineer H. Nyquist in 1928. In 1948, the founder of information theory, C. E. Shannon stated the theorem more clearly and formally defined it as “theorem”. Therefore, the sampling theorem is also called Nyquist theorem and Nyquist–Shannon theorem. The sampling theorem has many expressions, among them, the most basic expression is the sampling theorem in the time domain and the sampling theorem in the frequency domain. The sampling theorem has been widely used in the fields of digital telemetry systems, information processing, digital communication and sampling control theory.

In signal processing, sampling theorem is the bridge linking the continuous-time signal (analog signal) and discrete-time signal (digital signal). This theorem shows the relationship between the sampling frequency and the signal spectrum, and is the fundamental basis for the discretization of a continuous-time signal. It also gives the criteria for selecting an appropriate sampling rate that allows a discrete sampling sequence to capture all useful information from a continuous-time signal with a limited bandwidth.

1.2.1 A/D Conversion

The process of converting a continuous-time signal into its corresponding digital signal is called analog-to-digital (A/D) conversion, and vice versa is called a digital-to-analog (D/A) conversion process. They are necessary procedures for digital signal processing. Usually, the frequency band of the analog signal is relatively wide. Thus, before A/D conversion, an anti-aliasing filter is used to process the analog signal and make it a band-limited signal, which is then converted into a digital signal by A/D converter. Finally, the digital signal is sent to a digital signal analyzer or a computer to complete signal processing. According to different needs, the processed digital signal may be converted into an analog signal again through a D/A converter to drive the peripheral actuators.

A/D conversion includes sampling, quantization, coding and other processes. Its working principle is shown in Fig. 1.11.

- (1) **Sampling:** It is a process of using the periodic impulse train $p(t)$ to extract a series of discrete samples $x(nT_s)$ (where $n = 0, 1 \dots N$) from the continuous-time

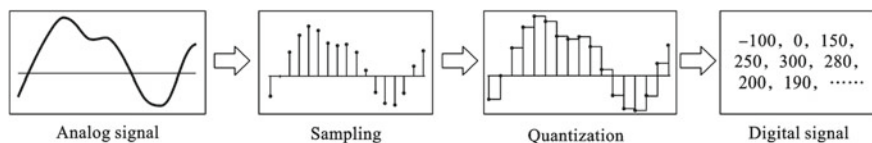
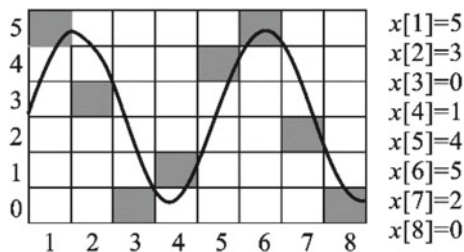


Fig. 1.11 Process of A/D conversion

Fig. 1.12 Process of quantization a signal



signal $x(t)$ to get a discrete sequence $x[n]$. In the following of this book, we will use round brackets “()” to indicate a continuous-time signal and square brackets “[]” to indicate a discrete-time signal. The periodic impulse train $p(t)$ is also called sampling function. T_s is called sampling period or sampling interval, and $f_s = 1/T_s$ is called the sampling frequency.

- (2) **Quantization:** It is a process of rounding and truncating the sampled signal $x(nT_s)$ to make it a number with limited number of significant digits. The signal value can only take certain numbers due to the limited bits of A/D converter. The higher the bits, the more values the discretized signal can take. In the quantization process illustrated in Fig. 1.12, the signal values can only take six different number from zero to five. Thus, the signal value is rounded to one of those numbers at each sampling point. It takes a period of time τ to finish each step of the quantization process, thus the instantaneous value of the analog signal should be constant to ensure the correctness and accuracy of the conversion. A sample-and-hold amplifier (SHA) is usually used to accomplish this goal.
- (3) **Coding:** It is a process of quantizing the discrete amplitude and turning it into a binary number, namely:

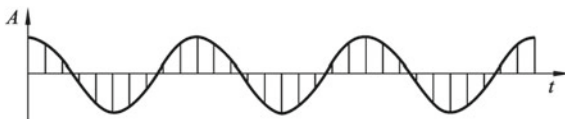
$$A = R \sum_{i=0}^{n-1} a_i 2^i \quad (1.6)$$

where, a_i is 0 or 1, R is the quantization increment or quantization step size. After the above transformation, the analog signal $x(t)$ becomes a digital signal that is discrete in time and quantized in amplitude. For the signal in Fig. 1.12, after coding with 4 bits, it becomes: $x[1] = 0101$, $x[2] = 0011$, $x[3] = 0000$, $x[4] = 0001$, $x[5] = 0100$, $x[6] = 0101$, $x[7] = 0010$, $x[8] = 0000$.

1.2.2 Sampling Error

The analog signal becomes a discrete signal with a limited number of data points after sampling. The interpolation between data points is approximated by a straight line, and the resulting error is called sampling error. The higher the sampling frequency, the smaller the sampling error.

Fig. 1.13 Sampling in practical engineering application



1.2.3 Nyquist–Shannon Sampling Theorem

In order to guarantee that the discretized digital signal can retain the main information of the original analog signal after sampling, the sampling frequency f_s must be at least twice the highest frequency component f_{\max} in the original signal. In practical engineering applications, the sampling frequency is at least 3 to 5 times the highest frequency component in the signal, as shown in Fig. 1.13.

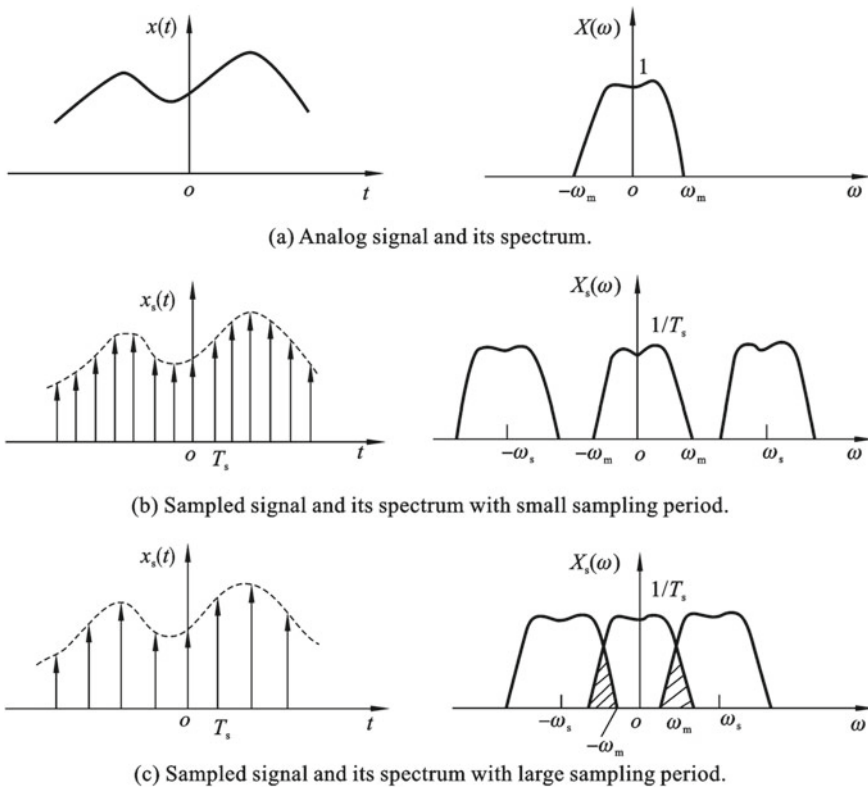
1.2.4 Frequency Aliasing in Sampling

The Nyquist–Shannon sampling theorem gives the minimum sampling frequency that allows undistorted sampling of the analog signal. If the sampling frequency does not meet the requirements, distortion will occur.

Aliasing is also called frequency aliasing effect. It is a phenomenon in which high and low frequency components are aliased (indistinguishable) after sampling. This effect happens when the sampling frequency is less than the one given by the sampling theorem. In this case, the high frequency components in the signal will be incorrectly sampled as low frequency components. Assume a signal $x(t)$ has frequency spectrum $X(\omega)$, and it is a band-limited signal with frequencies within the interval $(-\omega_m, \omega_m)$, as shown in Fig. 1.14. ω is called angular frequency, and it is related to frequency by the equation $\omega = 2\pi f$. The signal is sampled with sampling frequency ω_s , or sampling interval of $T_s = 2\pi/\omega_s$. When T_s is small, then $\omega_s > 2\omega_m$, and the periodic spectra are separated from each other as shown in Fig. 1.14b. When T_s is large, then $\omega_s < 2\omega_m$, and the periodic spectra are overlapped with each other as shown in Fig. 1.14c, this phenomenon is called aliasing. When aliasing happens, the high frequency information in the original signal is lost when the signal is sampled.

If a signal with frequency f_1 is sampled at f_2 , and the sampling frequency does not satisfy the sampling theorem, i.e. $f_2 < 2f_1$, then aliasing happens. The sampled signal will have a frequency of $f = |f_2 - f_1|$. As shown in Fig. 1.15a, b, when a signal with frequency of 1014 Hz is sampled at 1004 Hz and 1024 Hz, the sampled signal has the same waveform as the signal of 10 Hz. As shown in Fig. 1.15c, if the signals of 1014 Hz and 10 Hz are sampled at 4096 Hz, then the two signals can be clearly distinguished.

The MATLAB code for plotting Fig. 1.15a is given below:

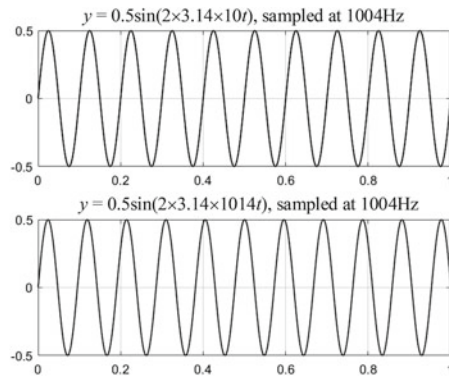
**Fig. 1.14** Effect of aliasing

```

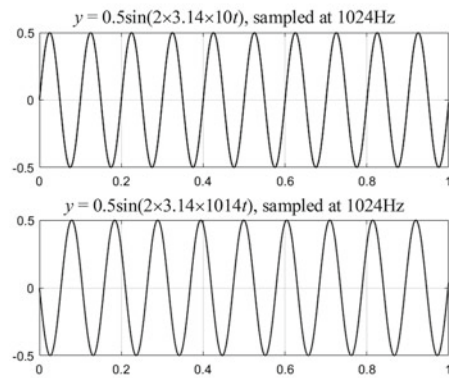
Fs = 1004;
dt = 1.0/Fs;
N = 1004;
T = N*dt;
t = linspace(0,T,N);
y1 = 0.5*sin(2*3.14*10*t);
figure
subplot(2,1,1)
plot(t,y1,'b','linewidth',1)
y2 = 0.5*sin(2*3.14*1014*t);
subplot(2,1,2)
plot(t,y2,'r','linewidth',1)

```

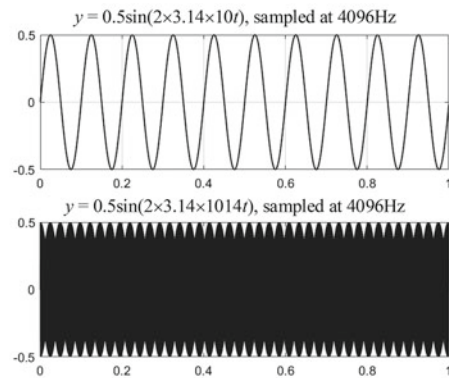
The MATLAB code for plotting Fig. 1.15b is given below:



(a) Sampled at 1004 Hz, which does not satisfy the sampling theorem and aliasing happens



(b) Sampled at 1024 Hz, which does not satisfy the sampling theorem and aliasing happens



(c) Sampled at 4096 Hz, which satisfies the sampling theorem and correct frequency information is obtained

Fig. 1.15 Aliasing effect on sampled signal

```

Fs = 1024;
dt = 1.0/Fs;
N = 1024;
T = N*dt;
t = linspace(0,T,N);
y1 = 0.5*sin(2*3.14*10*t);
figure
subplot(2,1,1)
plot(t,y1,'b','linewidth',1)
y2 = 0.5*sin(2*3.14*1014*t);
subplot(2,1,2)
plot(t,y2,'r','linewidth',1)

```

The MATLAB code for plotting Fig. 1.15c is given below:

```

Fs = 4096;
dt = 1.0/Fs;
N = 4096;
T = N*dt;
t = linspace(0,T,N);
y1 = 0.5*sin(2*3.14*10*t);
figure
subplot(2,1,1)
plot(t,y1,'b','linewidth',1)
y2 = 0.5*sin(2*3.14*1014*t);
subplot(2,1,2)
plot(t,y2,'r','linewidth',1)

```

1.2.5 Anti-Aliasing Filtering Before A/D Conversion

Anti-aliasing filter is a low-pass filter used before sampling. It can restrict the bandwidth of a signal to satisfy the sampling theorem. For a practical measurement signal, its frequency components that we concern are in a limited bandwidth. However, sometimes, there will be very high frequency noise in the signal. If the signal is sampled directly, then the high frequency component will be converted as a pseudo low frequency component that does not exist. To avoid this interference, an anti-aliasing filter can be used to remove the components with frequencies higher than $1/2$ of the sampling frequency. Thus, all the frequencies in the signal to be sampled satisfy the sampling theorem. Practically, the cutoff frequency f_c of the low-pass filter satisfies: $f_c = f_s/2.56$, where f_s is the sampling frequency of the A/D converter (Fig. 1.16).

Anti-aliasing is a necessary function for a measuring instrument to measure dynamic signals. The vibration signal of bridge, building and mechanical equipment has infinite bandwidth theoretically. Usually, we only concern the first few modes

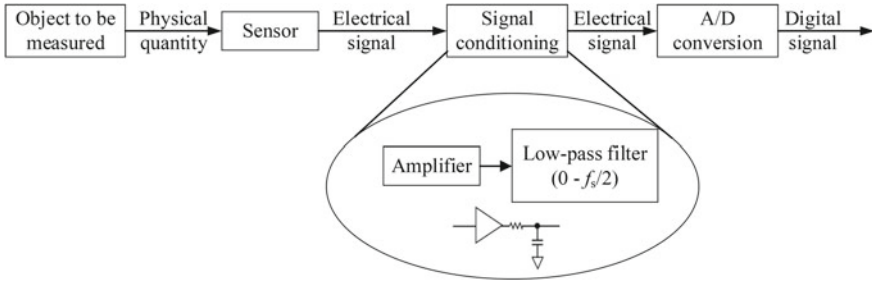


Fig. 1.16 Schematic diagram of anti-aliasing filtering before A/D conversion

that contribute most to the vibration. If we do not use anti-aliasing filter to remove the high frequency components, they may form false modes in the low frequency band, and make it difficult for us to analyze the vibration modes.

1.2.6 Quantization Error

The quantization of $x(nT_s)$ makes it a number with finite number of significant digits. The error introduced in this process is called quantization error. Take the largest value of $x(t)$ to be A , and divide it into D intervals, then each interval is $R = A/D$, where R is called quantization increment or quantization step. In the process of quantization, as shown in Fig. 1.12, $x(nT_s)$ should be rounded to integer multiplies R , which introduces error.

The quantization error has an equal probability distribution, and its probability density function is $p(x) = 1/R$. The maximum error introduced by rounding is $\pm 0.5R$, and the maximum error introduced by truncating is $-R$. The variance introduced by rounding is:

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 p(x) dx = \int_{-0.5R}^{0.5R} (x - \mu_x)^2 p(x) dx \quad (1.7)$$

Substituting $p(x) = 1/R$ and $\mu_x = 0$ into the equation, we get:

$$\sigma_x^2 = R^2/12$$

or

$$\sigma_x = 0.29R \quad (1.8)$$

The variance introduced by truncating can be calculated in the same way.

For example, if we use an N -bit A/D converter to sample and quantize an input voltage in the interval of (V_{\min}, V_{\max}) , the number of intervals is $D = 2^N - 1$, the

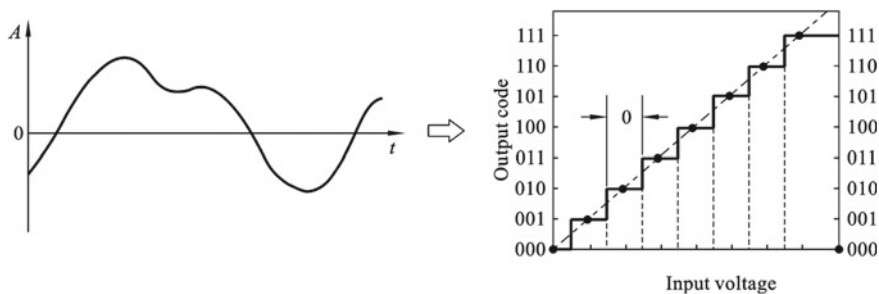


Fig. 1.17 Quantization in a 3-bit A/D converter

quantization increment is $R = (V_{\max} - V_{\min})/(2^N - 1)$. The maximum absolute quantization error e and maximum relative quantization error ε are respectively:

$$\begin{cases} e = \frac{V_{\max} - V_{\min}}{2(2^N - 1)} \\ \varepsilon = \frac{1}{2(2^N - 1)} \end{cases} \quad (1.9)$$

If truncating quantization is used, the maximum absolute quantization error e and maximum relative quantization error ε are respectively

$$\begin{cases} e = \frac{V_{\max} - V_{\min}}{2^N - 1} \\ \varepsilon = \frac{1}{2^N - 1} \end{cases} \quad (1.10)$$

Usually, the quantization error is considered as noises introduced in the process of A/D conversion, thus it is also called rounding noise or truncating noise. The quantization increment depends on the number of bits of the A/D conversion. The smaller the number of acquisition bits, the smaller the number of intervals D , and the larger the quantization increment R and quantization error. For example, the number of intervals of an 8-bit A/D converter is $D = 2^8 - 1 = 255$, the quantization increment is $R = V_{\max}/255$, and the quantization error introduced by rounding and truncating are $V_{\max}/510$ and $V_{\max}/255$ respectively (Fig. 1.17).

1.2.7 Technical Index of A/D Converter

The quality of an A/D converter mainly indicated by three indexes: resolution, conversion accuracy, and conversion speed.

1. Resolution

The resolution of an A/D converter is expressed by the number of bits of its output. The more the number of bits, the smaller the quantization increment, the smaller the quantization error, and the higher the resolution. Commonly used number of bits is 8-bit, 10-bit, 12-bit, 16-bit, 24-bit, 32-bit, etc. For an 8-bit A/D converter with input range of $-10\text{ V} - +10\text{ V}$, if we use the first bit to indicate the positive and negative signs and the other 7 bits to indicate amplitude, then the last bit is representing an analog voltage of 80 mV ($10\text{ V} \times 1/2^7 \approx 80\text{ mV}$). Namely, the minimum analog voltage that can be distinguished by the converter is 80 mV . In the same condition, the minimum analog voltage that can be resolved with a 10-bit converter is 20 mV ($10\text{ V} \times 1/2^9 \approx 20\text{ mV}$).

2. Conversion accuracy

If a converter with a certain resolution adopts the rounding method in the quantization process, its maximum quantization error should be half of the resolution value. As in the above example, the maximum quantization error of the 8-bit converter is 40 mV ($80\text{ mV} \times 0.5 = 40\text{ mV}$). It can be seen that the accuracy of an A/D converter is determined by the maximum quantization error. In fact, the last digits of many converters are not reliable, and the actual accuracy is even lower.

Since the analog-to-digital conversion module usually includes two parts: analog processing and digital conversion, the accuracy of the entire converter should also consider the errors of the analog processing parts (such as integrators, comparators, etc.). Generally, the analog processing error and the digital conversion error of the converter should be in the same order of magnitude, and the total error is the cumulative sum of these errors. For example, when a 10-bit A/D converter uses 9 bits to indicate amplitude, the maximum relative quantization error is $2^{-9} \times 0.5 \approx 0.1\%$. If the accuracy of the analog part can also reach 0.1% , the total accuracy of the converter is close to 0.2% .

3. Conversion speed

Conversion speed refers to the time it takes to complete a conversion, that is, the time it takes from sending the conversion control signal until the output terminal obtains a stable digital output. The longer the conversion time, the lower the conversion speed. The conversion speed is related to the conversion principle. For example, the conversion speed of the bit-by-bit approximation A/D converter is much higher than that of the double integral A/D converter. Besides, the conversion speed is also related to the number of bits of the converter. Generally, converters with a small number of bits (poor conversion accuracy) have a higher conversion speed. For the commonly used 8-bit, 10-bit, 12-bit, 16-bit converter, the conversion time is generally between a few microseconds to hundreds of milliseconds.

Since the converter must complete a conversion within the sampling interval T_s , the maximum signal frequency that the converter can handle is limited by the conversion speed. If a high-speed 10-bit A/D with conversion time of $50\text{ }\mu\text{s}$ is used, the sampling frequency can be as high as 20 kHz .

1.2.8 D/A Conversion

Conventional signals, such as sound, are analog signals, while the signals recorded in storage devices, such as the hard disk of a computer, are digital signals. Recording a song to the disk requires converting analog signals into digital signals, while playing the song from disk requires converting digital signals into analog signals and then playing them through a speaker. This process is accomplished by digital-to-analog (D/A) converters.

D/A conversion is the process of recovering an analog signal from a digital signal. It is generally implemented by a hold circuit, such as zero-order hold circuit and first-order multi-angle hold circuit. Zero-order hold circuit outputs a constant within a sampling period. The discontinuities (step-like) in the signal formed at the junction of two sampling periods must be smoothed with an analog low-pass filter. The process is shown in Fig. 1.18.

A D/A converter is a device that converts a digital signal into a voltage or current signal. D/A converter first converts the digital signal into an analog electrical impulse signal through a T-type resistor network, and then converts it into a stepped continuous electrical signal through a zero-order hold circuit. As long as the sampling interval is dense enough, the original signal can be reproduced accurately. In order to reduce the electrical noise caused by the zero-order hold circuit, a low-pass filter is connected in the end.

Generally speaking, compared with A/D converter, the circuit of D/A converter is relatively simple and the price is relatively low. Interestingly, many A/D converters include a D/A converter inside. There are many different types of circuit for D/A converters. The simplest is the weighted resistor network D/A converter as shown in Fig. 1.19. There are also other D/A converters developed based on it, such as two-stage weighted resistance network D/A converters, inverted T-type resistance network D/A converters, and bipolar output voltage D/A converters. The main technical indexes of D/A converters are also resolution, conversion accuracy and conversion speed.

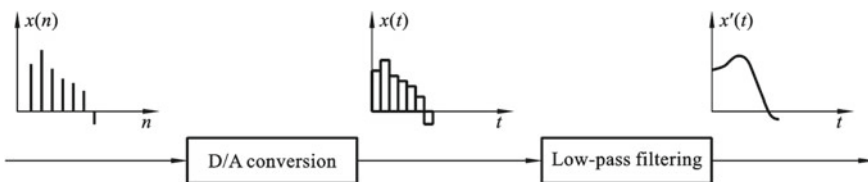


Fig. 1.18 Process of D/A conversion

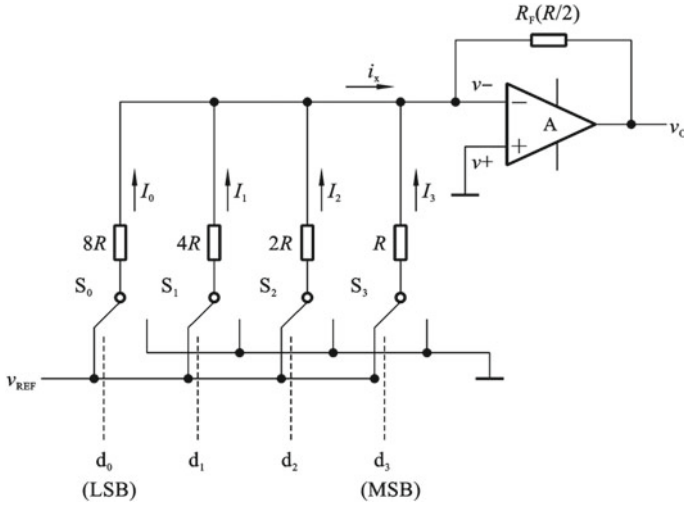


Fig. 1.19 Weighted resistor network D/A converter

1.3 Standard Functions in Signal Analysis

In measurement and signal analysis, standard functions play an important role. They can be used to analyze the performance of the measurement system. The standard functions mainly include: delta function, unit step function, unit ramp function, exponential function, sine function, sinc function, white noise, etc.

1.3.1 Unit Impulse Function (Δ Function)

1. δ function

δ function is also called delta function, unit impulse function or Dirac function. It was first used by Paul Dirac in quantum mechanics in 1930. An ideal δ function is zero everywhere except at the origin, which is infinitely high. However, the area of the pulse is limited, equal to one. The change process is shown in Fig. 1.20. The shadow area is kept as 1, as the time axis gets narrower, its height gets higher and higher. In an ideal situation, the time duration approaches zero while the height extends to infinity, then it becomes δ function.

Mathematically, the process can be described as follows. A square function $S_\varepsilon(t)$ with area of 1 and duration of ε can be expressed as:

$$S_\varepsilon(t) = \begin{cases} 1/\varepsilon, & 0 \leq t \leq \varepsilon \\ 0, & t < 0, t > \varepsilon \end{cases} \quad (1.11)$$

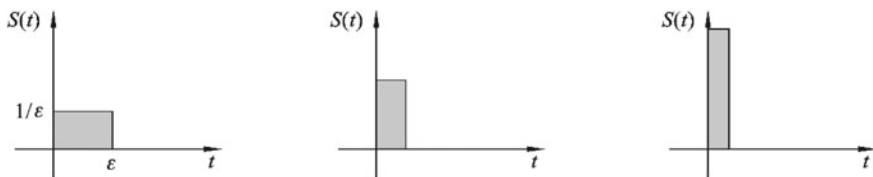


Fig. 1.20 Ideal unit impulse function

When ϵ becomes smaller, its value $1/\epsilon$ becomes larger. And when ϵ approaches zero ($\epsilon \rightarrow 0$), the function value tends to infinity:

$$\delta(t) = \begin{cases} \infty, & t=0 \\ 0, & t \neq 0 \end{cases} \quad (1.12)$$

As for the area, it is:

$$\int_{-\infty}^{\infty} \delta(t) dt = \lim_{\epsilon \rightarrow 0} \int_{-\infty}^{\infty} S_{\epsilon}(t) dt = 1 \quad (1.13)$$

From a physical point of view, the δ function is an ideal function and a physically unrealizable signal. Because, when a signal is generated by any instrument, its duration can never be zero, and the amplitude can never be infinite.

2. Properties of Δ Function

(1) Multiplication (sampling):

$$f(t)\delta(t) = f(0)\delta(t) \quad (1.14)$$

or

$$f(t)\delta(t-t_0) = f(t_0)\delta(t-t_0) \quad (1.15)$$

(2) Integration (sifting):

$$\int_{-\infty}^{\infty} f(t)\delta(t) dt = f(0) \int_{-\infty}^{\infty} \delta(t) dt = f(0) \quad (1.16)$$

or

$$\int_{-\infty}^{\infty} f(t)\delta(t-t_0) dt = f(t_0) \int_{-\infty}^{\infty} \delta(t-t_0) dt = f(t_0) \quad (1.17)$$

(3) Convolution:

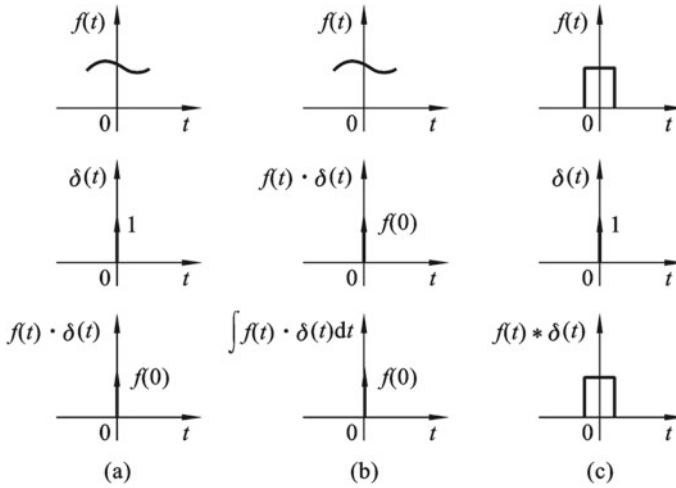


Fig. 1.21 Properties of δ function

$$f(t) * \delta(t) = \int_{-\infty}^{\infty} f(\tau) \delta(t-\tau) d\tau = f(t) \quad (1.18)$$

The above properties of δ function can be graphically shown in Fig. 1.21.

3. Transforms of Δ Function

(1) Laplace transform

$$\Delta(s) = \int_{-\infty}^{\infty} \delta(t) e^{-st} dt = 1 \quad (1.19)$$

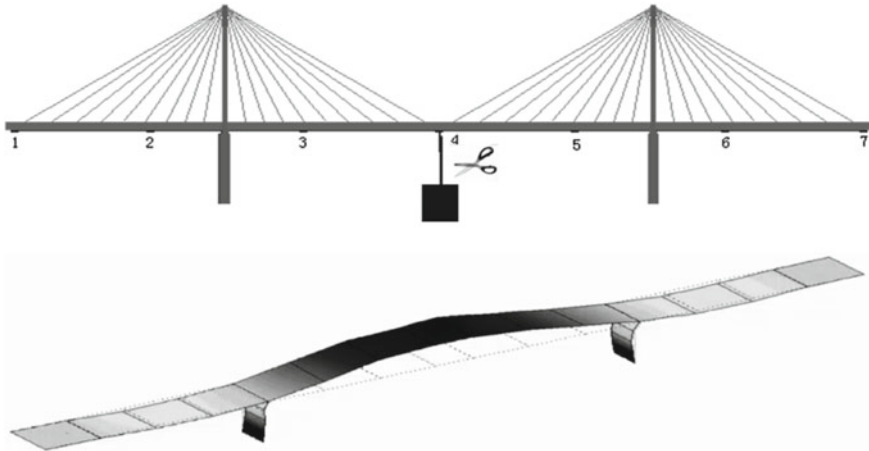
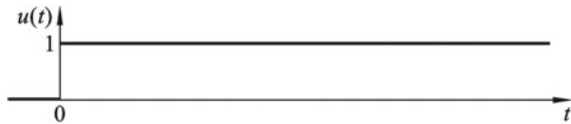
(2) Fourier transform

$$\Delta(f) = \int_{-\infty}^{\infty} \delta(t) e^{-j2\pi ft} dt = 1 \quad (1.20)$$

1.3.2 Unit Step Function

Unit step function was proposed by Oliver Heaviside, thus it is also called Heaviside step function. As shown in Fig. 1.22, unit step function is zero for $t < 0$, and constantly one for $t \geq 0$. It is a discontinuous function.

$$u(t) = \begin{cases} 0, & t < 0 \\ 1, & t \geq 0 \end{cases} \quad (1.21)$$

Fig. 1.22 Unit step function**Fig. 1.23** Measurement of natural frequency of bridge

The unit step function and unit impulse function can relate each other by integration and differentiation:

$$\begin{cases} u(t) = \int \delta(t) dt \\ \delta(t) = \frac{du(t)}{dt} \end{cases} \quad (1.22)$$

In the analysis of bridge bending, equation for bending moment is used. The commonly used bending moment equation is expressed by a piecewise function, which brings inconvenience to theoretical analysis. Through the use of unit step function, the bending moment equation expression can be expressed as a global equation, which greatly simplifies the calculation of the deformation. For example, the unit step function is used for analyzing the natural frequency of bridge as shown in Fig. 1.23.

1.3.3 Unit Ramp Function

A unit ramp function is shown in Fig. 1.24. Its value is zero for $t < 0$, and is proportional to the independent variable for $t \geq 0$.

Fig. 1.24 Unit ramp function



$$v(t) = \begin{cases} 0, & t < 0 \\ t, & t \geq 0 \end{cases} \quad (1.23)$$

The unit ramp function is related to unit step function by integration and differentiation, as shown in Eq. (1.24).

$$\begin{cases} v(t) = \int u(t) dt \\ u(t) = \frac{dv(t)}{dt} \end{cases} \quad (1.24)$$

1.3.4 Complex Exponential Function

Complex exponential function is expressed as e^{st} ($-\infty, +\infty$), where $s = \sigma + j\omega$ is a complex number. By choosing different forms of s , the exponential function can represent various waveforms, as shown in Fig. 1.25. In the figure, the horizontal axis and vertical axis represent real and imaginary parts of s respectively.

- (1) When s is a real number, i.e. $\omega = 0$, if $\sigma \neq 0$, then it represents an ascending or descending exponential function; if $\sigma = 0$, then it represents a DC signal with constant value.
- (2) When s is a pure imaginary number, i.e. $\sigma = 0$, if $\omega \neq 0$, then

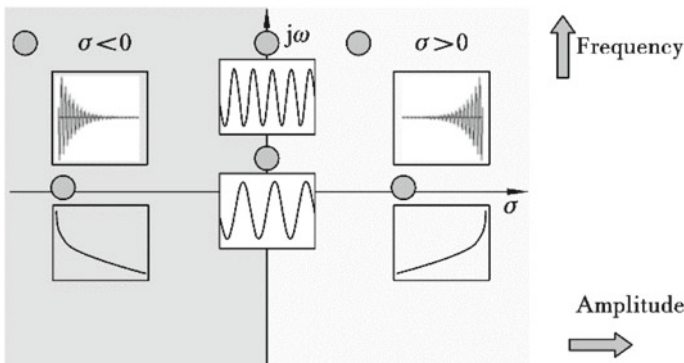


Fig. 1.25 Correspondence between s and signal waveform

$$e^{st} = e^{j\omega t} = \cos \omega t + j \cdot \sin \omega t \quad (1.25)$$

in which, the real part $\text{Re}(e^{j\omega t}) = \cos \omega t$ represents a cosine function and its imaginary part $\text{Im}(e^{j\omega t}) = \sin \omega t$ represents a sine function.

- (3) When s is a complex number, i.e. $\sigma \neq 0$ and $\omega \neq 0$, then

$$e^{st} = e^{\sigma t} \cdot e^{j\omega t} = e^{\sigma t} \cos \omega t + e^{\sigma t} j \sin \omega t \quad (1.26)$$

$\text{Re}(e^{st}) = e^{\sigma t} \cos \omega t$ represents an ascending or descending cosine function and its imaginary part $\text{Im}(e^{st}) = e^{\sigma t} \sin \omega t$

Expressing the above mentioned situations in the s -plane, it can be seen that each point in the s -plane corresponds to a certain exponential function mode as shown in Fig. 1.25. Its vertical axis $j\omega$ represents the oscillation of e^{st} while its horizontal axis σ represents the amplitude change of e^{st} . When $\omega = 0$ and s changes along the horizontal axis from left to right, the signal waveform changes from a descending exponential function ($\sigma < 0$) to a constant ($\sigma = 0$), then to an ascending exponential function ($\sigma > 0$); When $\sigma = 0$ and s changes along the positive half of the vertical axis, the signal waveform is sinusoidal function with increasing frequency. One may ask why the negative half of the vertical axis is not considered. The negative half of the vertical axis is symmetrical to the positive half with a change of the sign. In the negative half, the frequencies are negative. As we already learned, frequency represents the number of repetitions per unit time. Thus the negative frequency does not have a real physical meaning and only comes from the mathematical calculations.

The complex exponential function also has some important properties as follows:

- (1) Any time-dependent function can be expressed as the discrete sum or continuous integral of complex exponential functions. These are so-called Fourier series and Fourier transform that we are going to learn in following chapters.

In discrete form:

$$x[n] = \sum_r C_r e^{sn} \quad (1.27)$$

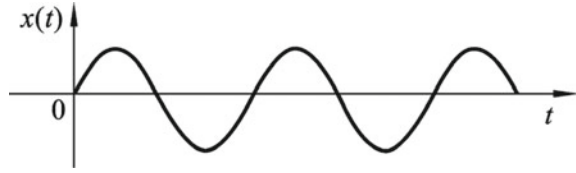
In continuous form:

$$x(t) = \int_{S_A}^{S_B} C_s e^{st} dt \quad (1.28)$$

- (2) The complex exponential e^{st} always exists in its analyzing function when it is differentiated, integrated or passed through a linear system.

When it is differentiated:

$$\frac{d}{dt} e^{st} = s e^{st} \quad (1.29)$$

Fig. 1.26 Sine function

When it is integrated:

$$\int e^{st} dt = \frac{e^{st}}{s} \quad (1.30)$$

When it is passed through a linear system

$$e^{st} \xrightarrow{H} H(s)e^{st} \quad (1.31)$$

where $H(s)$ is the system response function. These properties show the perpetual nature of the e^{st} function.

1.3.5 Sine Function

The name of the sine function comes from the sinusoidal curves. It is very common in science and mathematics and can be expressed as:

$$x(t) = A \sin(2\pi ft + \varphi) \quad (1.32)$$

Many signals are sine functions, such as the electromagnetic waves and alternating current (Fig. 1.26).

1.3.6 Sinc Function

Sinc function is also called gate function, filter function or interpolation function. Its name is abbreviated from “sine cardinal”. It frequently appears in signal analysis attributed to its simple Fourier transform result, i.e. a rectangular function. Sinc function is defined as:

$$\text{sinc}(t) = \frac{\sin(t)}{t} \quad (-\infty < t < \infty)$$

or

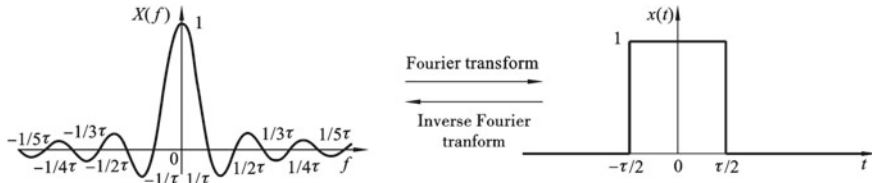


Fig. 1.27 Fourier transform of sinc function

$$\text{sinc}(t) = \frac{\sin(\pi t)}{\pi t} \quad (-\infty < t < \infty) \quad (1.33)$$

It is an even function, which decays gradually in the positive and negative directions of t axis, as shown in Fig. 1.27. When $t = \pm\pi, \pm2\pi, \dots, \pm n\pi$, the function value is zero, and when $t = 0$, the function value is 1. The sinc function and rectangular function are a set of Fourier transform pairs. It is called a filter function because it can achieve low-pass filtering when any signal is convolved with the $\text{sinc}(t)$ function in time domain; the reason why it is called an interpolation function is because when the sampled signal is restored, it is superimposed by several $\sin(t)$ functions in the time domain to form a non-sampling point waveform.

1.3.7 White Noise

White noise is a random signal with a flat power spectral density function. That is, for all frequencies f , we have:

$$S_x(f) = \frac{N_0}{2} \quad (1.34)$$

And the mean is zero:

$$\mu_x = \frac{1}{T} \int_0^T x(t) dt = 0 \quad (1.35)$$

A white noise signal is illustrated in Fig. 1.28.

In health care applications, white noise is used to treat people with sensitive hearing, such as shielding background noise to assist sleep.

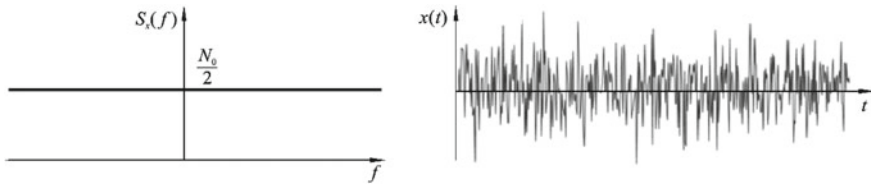


Fig. 1.28 White noise signal

1.4 Generation of Standard Functions

Signal generator is an electronic device that can generate a large number of standard signals and user-defined signals, with high precision, high stability, repeatability and easy operation. The signal generator has the advantages of continuous phase conversion and frequency stability. It can not only simulate various complex signals, but also dynamically and timely control the frequency, amplitude, phase shift, and waveform. It can also communicate with other instruments in an automatic measurement and test system, so it is widely used in the fields of automatic control systems, vibration excitation, communications, and instrumentation.

The signals that can generated include sine waves, square waves, sawtooth waves and white noise. The main components of the signal generator include a frequency generating unit, a modulation unit, a buffer amplifying unit, an attenuation output unit, a display unit, and a control unit. Early signal generators used analog circuits. With the development of phase-locked loop (PLL) frequency synthesizer, modern signal generators are commonly controlled by digital circuits or microcomputers. At present, the high-end signal generator adopts more advanced DDS frequency direct synthesis technology, which has the advantages of high frequency output stability, wide frequency synthesis range, and high signal spectrum purity. Signal generators can be divided into ultra-low frequency signal generators, low frequency signal generators, high frequency signal generators, and microwave signal generators according to the working frequency band. A typical signal generator is shown in Fig. 1.29.

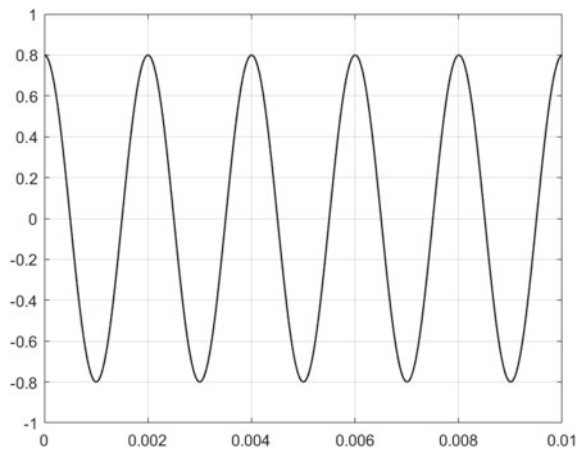
MATLAB provides codes for generating standard functions: “sin” for sine function, “square” for square function, “sawtooth” for sawtooth function, “randn” for white noise, “sinc” for sinc function, “pulstran” for pulse sequence, “gauspuls” for Gaussian sine pulse signal and “chirp” for chirp function. Examples of generating standard functions in MATLAB are given below.

Example 1.4 Sine Function Generator The MATLAB code for generating a sine function with amplitude of 0.8, frequency of 500 Hz and phase of 90° is shown below (Fig. 1.30).



Fig. 1.29 Signal generator

Fig. 1.30 Sine wave generated in MATLAB



```

Fs = 44,100; % Sampling frequency
dt = 1.0/Fs; % Sampling period
T = 1; % Duration of the signal
N = T/dt; % Number of sampling points
x = linspace(0,T,N); % Generate sampling points
A = 0.8; F = 500; P = 90; % Define amplitude, frequency and phase
y = A*sin(2*3.14*Fs*x + P*pi/180.0); % Generate signal
plot(x,y,'b','linewidth',1) % Plot the curve
xlim([0,0.01]) % Set range for x-axis
ylim([-1,1]) % Set range for y-axis
grid on % Turn on grids

```

Example 1.5 Square Function Generator The MATLAB code for generating square function with amplitude of 0.8, frequency of 500 Hz and duty cycle of 50% is shown below (Fig. 1.31).

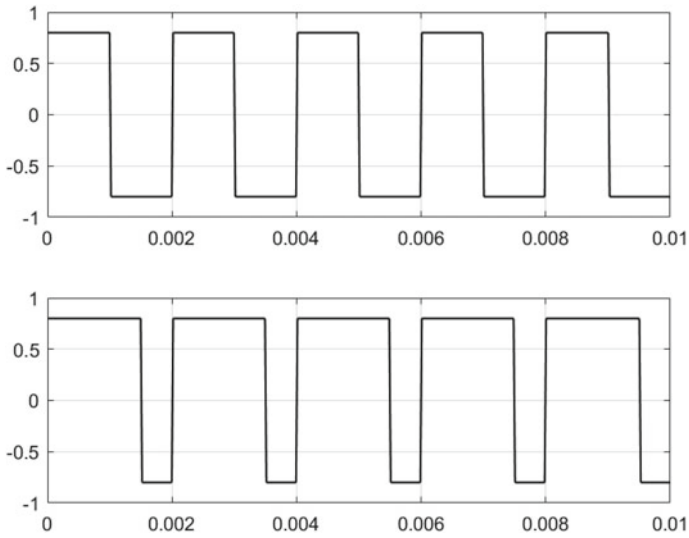


Fig. 1.31 Square wave generated in MATLAB

```

Fs = 44,100;
dt = 1.0/Fs;
T = 0.1;
N = T/dt;
x = linspace(0,T,N);
y = 0.8*square(2*3.14*500*x);
subplot(2,1,1);
plot(x,y,'b','linewidth',1);
xlim([0,0.01]);
ylim([-1,1]);
grid on;
y1 = 0.8*square(2*3.14*500*x,75);
subplot(2,1,2);
plot(x,y1,'b','linewidth',1);
xlim([0,0.01]);
ylim([-1,1]);
grid on;

```

Example 1.6 Sawtooth Function Generator The MATLAB code for generating sawtooth function with amplitude of 0.8 and frequency of 500 Hz is shown below (Fig. 1.32).

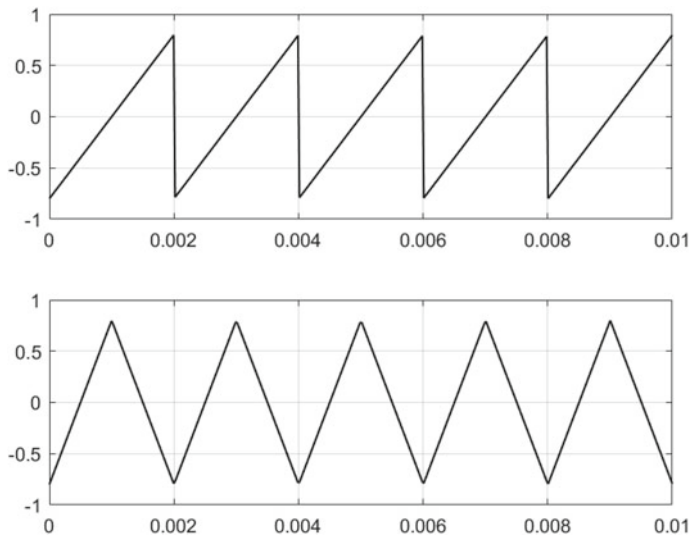


Fig. 1.32 Sawtooth wave generated in MATLAB

```

Fs = 44,100;
dt = 1.0/Fs;
T = 0.1;
N = T/dt;
x = linspace(0,T,N);
y = 0.8*sawtooth(2*3.14*500*x);
subplot(2,1,1);
plot(x,y,'b','linewidth',1);
xlim([0,0.01]);
ylim([-1,1]);
grid on;
y1 = 0.8*sawtooth(2*3.14*500*x,0.5);
subplot(2,1,2);
plot(x,y1,'b','linewidth',1);
xlim([0,0.01]);
ylim([-1,1]);
grid on;

```

Example 1.7 White Noise Generator The MATLAB code for generating a white noise is shown below (Fig. 1.33).

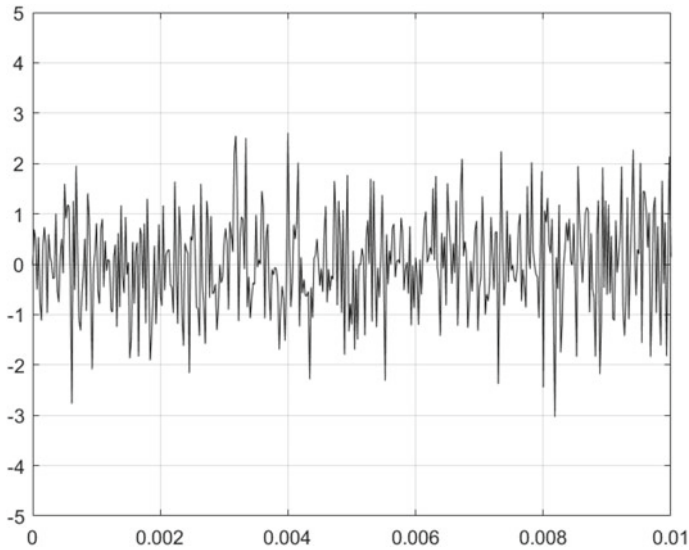


Fig. 1.33 White noise generated in MATLAB

```

Fs = 44,100;
dt = 1.0/Fs;
T = 0.01;
N = T/dt;
x = linspace(0,T,N);
y = randn(1,N);
plot(x,y,'b','linewidth',0.5);
xlim([0,0.01]);
ylim([-5,5]);
grid on;

```

1.5 Waveform Analysis of Signals

Measurement signal is usually regarded as a quantity that varies with time, namely $x(t)$. A curve drawn with the signal amplitude as ordinate and time as abscissa is called the signal waveform, as shown in Fig. 1.34. The most natural and direct method in signal analysis is waveform analysis. Because the analysis is carried out in the time domain, it is also called time domain analysis. Time domain analysis is the most intuitive analysis, from which some primitive judgments can be made.

Fig. 1.34 Signal waveform



Many important parameters of the signal can be obtained in time domain analysis, such as the instantaneous value of the signal at any time, the maximum and minimum of the signal, the period and initial phase of periodic signals, the mean and variance of the signal. In addition, we can also observe starting time, duration, time lag, and waveform distortion through time-domain waveform analysis. Oscilloscope is one of the most commonly used waveform analysis tools.

Important parameters in waveform analysis include: period, initial phase, peak value, mean value, mean square value, and variance.

1. Signal Period

For periodic signals, the signal period refers to the time required for the signal amplitude to complete a periodic change. For example, a sine wave signal $x(t) = A\sin(2\pi f_0 t)$ is a periodic signal whose period is the reciprocal of frequency f_0 , i.e. $T = 1/f_0$.

2. Initial Phase

The initial phase of a periodic signal refers to the position of the signal in the cycle when time is zero. It is usually measured in the unit of degrees (angle), thus it is also called the initial phase angle. For example, a sine wave signal $x(t) = A\sin(2\pi f_0 t + \varphi_0)$ has the initial phase of φ_0 .

3. Signal Amplitude

The peak value of the signal includes positive peak value P_p (maximum value), negative peak value P_{-p} (minimum value) and peak-to-peak value P_{p-p} (maximum value minus minimum value) of the signal, as shown in Fig. 1.35.

Performing statistical calculations on the time-domain signal, the first-order statistics (e.g. mean), second-order statistics (e.g. mean square value, variance, correlation, power spectrum, etc.) as well as higher-order statistics (e.g. slope) of the measured signal can be obtained.

4. Mean

The mean value of a signal $E[x(t)]$ represents the mean value of a set or the mathematical expectation value. Based on the ergodic nature of the stochastic process, it can be represented by the average value of the amplitude in a time interval T , namely

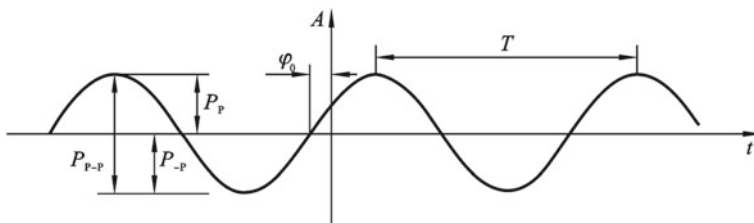


Fig. 1.35 Parameters of in waveform analysis

$$\mu_x = E[x(t)] = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) dt \quad (1.36)$$

The mean value $E[x(t)]$ reflects the change trend of the center of the measured signal, and it is also called DC value (Fig. 1.36).

5. Mean Square Value

The mean square value of a signal $E[x^2(t)]$, or average power, indicates the strength of the signal.

$$\psi_x^2 = E[x^2(t)] = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x^2(t) dt \quad (1.37)$$

The root mean square (RMS) value of a signal (i.e. the positive square root of the mean square value) is also called the effective value, and it is also a common indication for the average strength and energy of a signal. In engineering measurement, mean square value is used to identify local abnormalities, such as a breakage in wire rope.

6. Variance

Variance of a signal $x(t)$ is defined as:

$$\sigma_x^2 = E[(x(t) - E[x(t)])^2] = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T (x(t) - \mu_x)^2 dt \quad (1.38)$$

It reflects the degree of fluctuation of the signal around the mean value. For example, as shown in Fig. 1.37, variance is used for surface roughness evaluation, in which large variance reflects large disturbances and rougher surface. σ_x is called the mean square error or standard deviation.

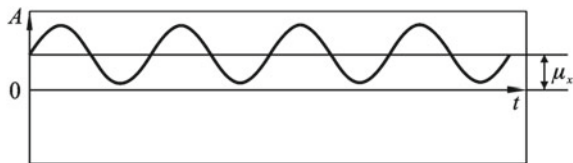
σ_x^2 describes the fluctuation amount of the signal, while μ_x describes the DC component of the signal. Signals with different variance and mean value are shown in Fig. 1.38.

It can be easily proved that the following relationship exists between the mean, mean square and variance of the signal:

$$\psi_x^2 = \sigma_x^2 + \mu_x^2 \quad (1.39)$$

Example 1.8 Fault Diagnosis of Bearing For engineers, waveform analysis is an ideal tool that can be used to diagnose a series of machine failures, including bearing

Fig. 1.36 Mean value of a signal



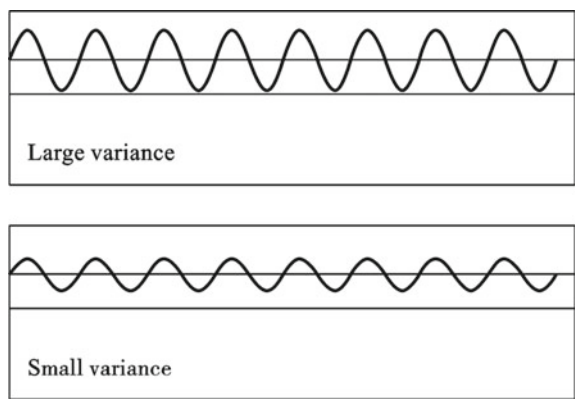


Fig. 1.37 Signal variance

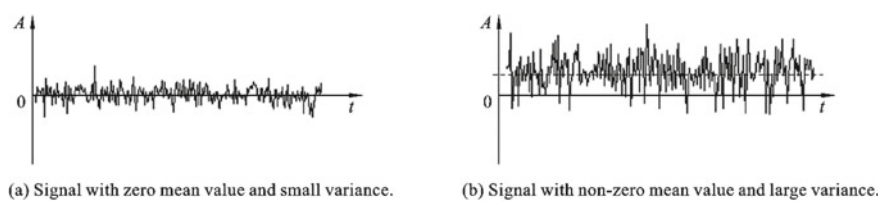


Fig. 1.38 Signals with different variance and mean value

failure, gear failure, cavitation, friction, looseness, etc. As shown in Fig. 1.39, the sudden change of signal amplitude (local change of variance) indicates the existence of a crack in the bearing.

Example 1.9 Waveform Analysis of bird Sounds Cuckoos like to sing during breeding, they often stand on the top branches of trees to sing endlessly. Sometimes

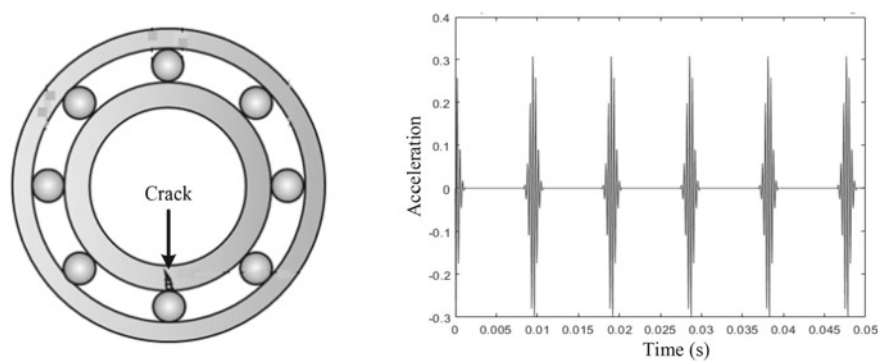


Fig. 1.39 Bearing fault diagnose based on waveform analysis

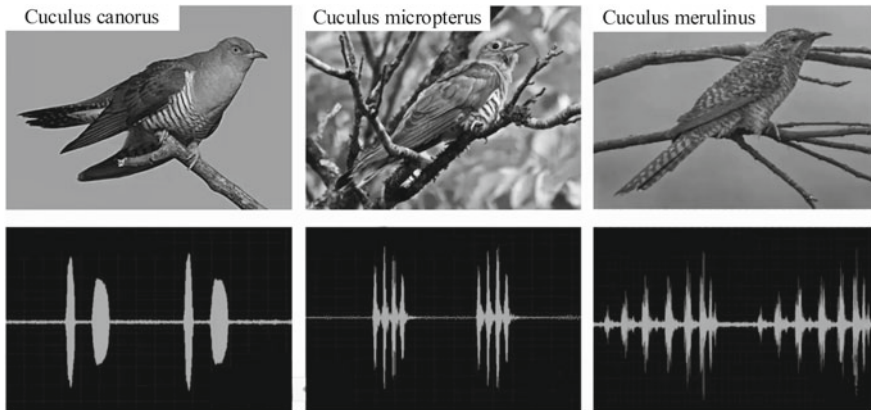


Fig. 1.40 Sound waveforms of different cuckoos

it sings at night or while flying. The loud call of “Cuckoo ~ Cuckoo ~” can be heard from a long distance away. The call usually repeats 20 times per minute. There are many types of cuckoos that sing differently. Thus the types of cuckoos can be distinguished through waveform analysis of their calls. The sound signal can be obtained through a microphone. The period, root mean square, peak frequency and other characteristics of the signal can be extracted, and then the signal can be identified and classified by pattern recognition, as shown in Fig. 1.40.

Exercise

1. Briefly describe the classifications of signals and the characteristics each type.
2. What are the differences between signal analysis methods for deterministic signal and non-deterministic signal?
3. Explain aliasing, leakage and fence effects in the process of signal discretization, and show how to prevent these phenomena from occurring.
4. What is the sampling theorem? What is the Nyquist frequency? How to avoid frequency domain aliasing and distortion of the sampled signal?
5. Find the mean value μ_x , mean square value ψ_x^2 and probability density function $p(x)$ of the sinusoidal signal $x(t) = A \sin(\omega t + \phi)$.
6. Find the mean value and mean square value of the sine signal $x(t) = \sin 200t$.
7. Sampling the 500 Hz sine signal $x(t) = \sin(2\pi \cdot 500 \cdot t)$ with sampling frequency $f_s = 4096$ Hz and total number of sampling points $N = 2048$. Find the Nyquist frequency and the frequency resolution of the sampling signal. If the signal contains several harmonic components of the fundamental frequency of 500 Hz, which harmonics can be accurately detected? Which harmonics will be distorted in the sampling?
8. To sample a band-limited signal with the highest frequency of 2 kHz, the frequency resolution is required to be 1 Hz. How to choose the sampling frequency and the number of sampling points?

9. Sampling three sine signals $x_1(t) = \sin 2\pi t$, $x_2(t) = \sin 6\pi t$ and $x_3(t) = \sin 10\pi t$ with sampling frequency of $f_s = 4$ Hz. Find the respective output sequences, draw the time-domain waveforms, analyze the sampling results of the three signals, and explain the phenomenon of frequency aliasing.

Chapter 2

Frequency Domain Analysis



Frequency domain analysis is to rearranged a signal with its frequency as the abscissa. Frequency domain analysis is also called spectrum analysis. Furthermore, according to the different dependent variables, frequency spectrum is subdivided into: magnitude spectrum, phase spectrum, power spectrum, amplitude density spectrum, energy density spectrum, power density spectrum, etc. Besides, the types of spectrum are also dependent on the signals to be analyzed, as shown in Fig. 2.1.

The continuous spectrum analysis method is used for the analog signal, and the discrete spectrum analysis method is used for the digital signal. The mathematical tools of the former are Fourier Transform (FT) and Fourier Series (FS), and the latter is Discrete Fourier Transform (DFT).

2.1 Concept of Frequency Domain Analysis

For any signal that can be represented by time-varying amplitude, then there is a corresponding frequency spectrum. Visible light, music, radio waves and vibration all have such properties. Spectrum was first used in the field of optics to describe the rainbow colors separated by prisms, as shown in Fig. 2.2. The sun light is composed of different colors, and the light of each color has a different frequency. Prisms refract light of different frequencies to different positions on a screen through refraction, so different colors of light can be seen. The color of each light band indicates its frequency, and the brightness indicates the intensity of light. This is the spectrum of light.

The sound emitted by a sound source may also compose sounds of many different frequencies. Different frequencies will stimulate their corresponding receivers in the ears. If the main stimulus has only one frequency, its pitch can be heard. A harmonic sound is composed of a fundamental frequency and its multiples. The timbre of the sound is determined by the high frequency parts, which are called overtone. A noisy sound usually contains various frequencies without multiplication relationships, its

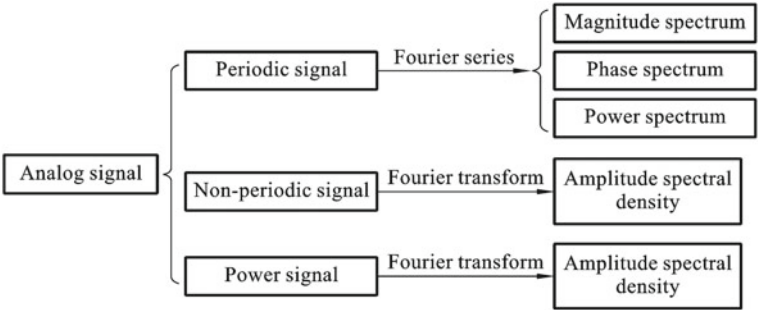


Fig. 2.1 Spectra for different types of signals

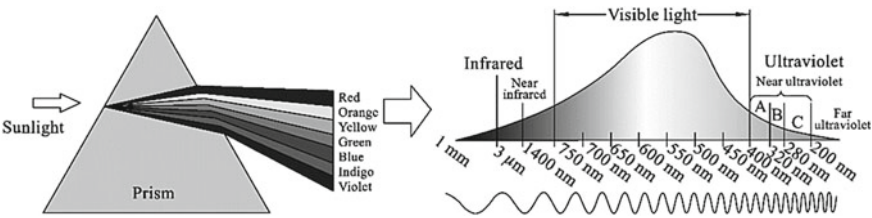


Fig. 2.2 Spectrum of sun light

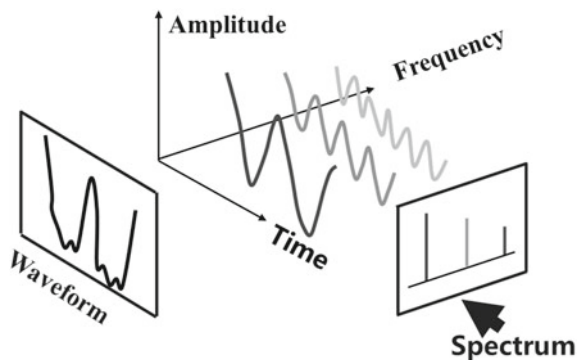
spectrum is usually a horizontal line. Usually, the sound can be heard by humans is between 20 Hz to 20 kHz. The sound with frequency lower than 20 Hz is called infrasound, and the one with frequency higher than 20 kHz is called ultrasound.

The frequency spectrum specifically refers to the representation of a time domain signal in the frequency domain. Many physical signals can be expressed as the sum of simple signals of different frequencies. Frequency domain analysis is to find the information (amplitude, phase or power, etc.) of a signal at different frequencies. Specifically, Fourier series expansion or Fourier transform can be performed on the signal, thereby decomposing the signal into sine and cosine signals with different frequencies. The result can be expressed with amplitude or phase as the vertical axis and frequency as the horizontal axis.

As shown in Fig. 2.3, time domain analysis and frequency domain analysis are two perspectives of viewing a signal. Generally, the time domain analysis of a signal is the most intuitive one, especially for simple periodic signals. When analyzing a complex signal containing many frequency components, such as the one shown in Fig. 2.3, time domain analysis is no longer easy to identify the information of the signal. While observing from the frequency domain, it can be clearly seen that this signal has three frequency components. The tool used to switch from time domain to frequency domain is the Fourier transform.

The frequency spectrum uses frequency as the abscissa and the amplitude of each frequency as the ordinate and to indicate the frequency composition of the measured signal. Waveform analysis only reflects the change of signal amplitude over time, it

Fig. 2.3 The difference between frequency domain analysis and time domain analysis



is difficult to observe the frequency composition of the signal, while the frequency spectrum can directly display the frequency composition.

2.1.1 Advantages of Frequency Domain Analysis

- (1) Frequency domain analysis has a clear physical meaning. When physical phenomena such as visible light, music, radio waves and vibration are analyzed in frequency domain, the cause of the signal and its related information can be explored. For example, the vibration frequency of a tuning fork can be analyzed as shown in Fig. 2.4; the vibration of a gearbox can be analyzed in frequency domain as shown in Fig. 2.5, from which, we can further analyze the transmission and meshing of gears.
- (2) Complex signal analysis. Simple sinusoids are easier to identify from the waveform, but the time domain waveforms for signals involving multiple harmonics

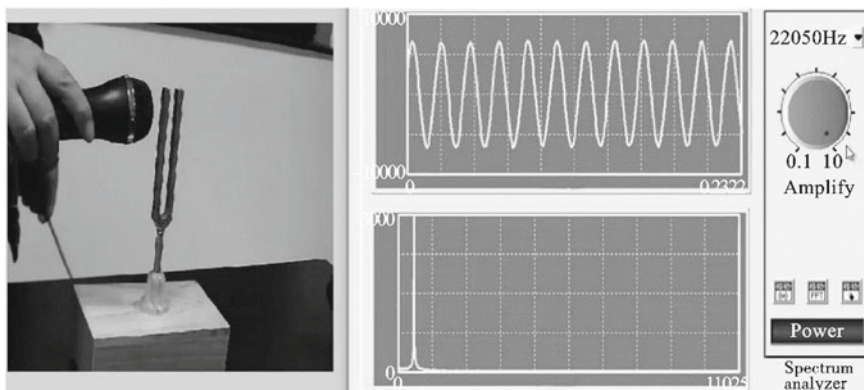


Fig. 2.4 Tuning fork and its frequency spectrum

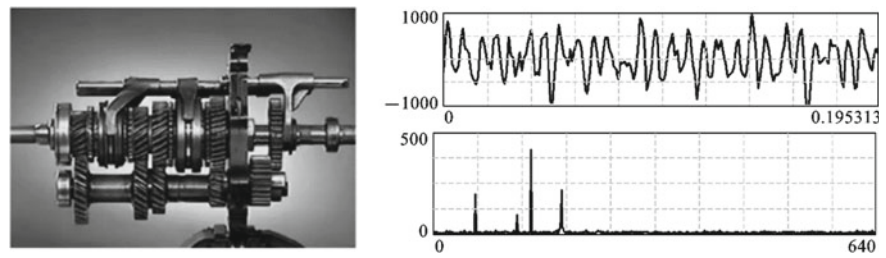


Fig. 2.5 Gearbox vibration signal and its frequency spectrum

or other components are complicated. For those signals, time domain waveform analysis is very difficult to implement. As shown in Fig. 2.6, a simple sinusoidal signal has one frequency component, and it is easy to be observed in time domain. Whereas, for the complex signals with multiple frequencies, their frequency components are difficult to be found in time domain while easy to be seen in frequency domain.

(3) Strong anti-noise ability. Actual signals usually contain noises. In the time domain, the signal is not smooth and has lots of spikes. It increases the difficulty of information identification. In the frequency domain, the noise is usually weak and easy to be separated, and main components in signal are easy to be highlighted, as shown in Fig. 2.7.

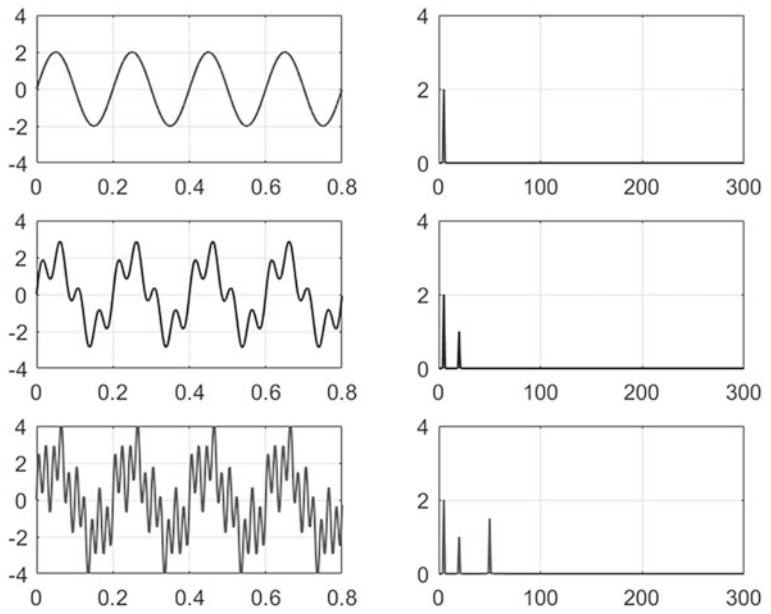


Fig. 2.6 Time domain waveform and spectrum of signals

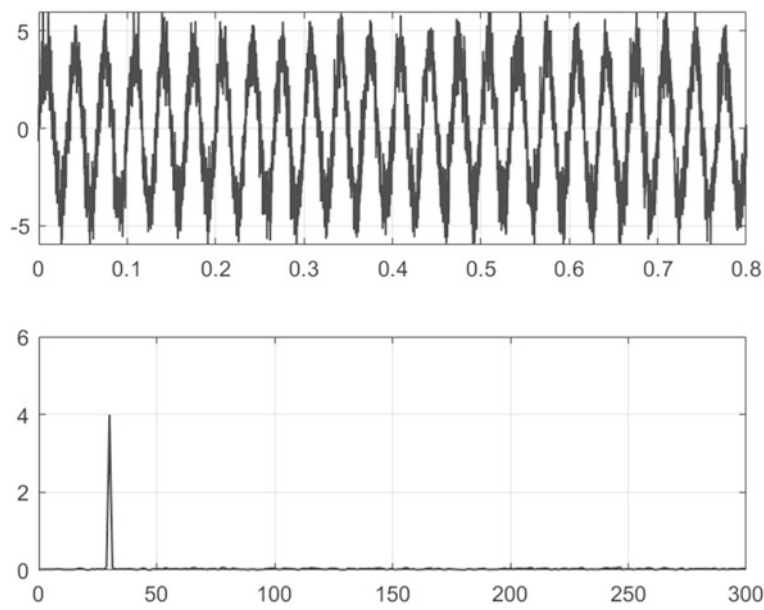


Fig. 2.7 Time domain waveform and frequency spectrum of noisy signal

2.1.2 Objectives of Frequency Domain Analysis

Frequency domain analysis is one of the most commonly used signal analysis methods. It uses Fourier transform or Fourier series to convert a time domain signal $x(t)$ into a frequency domain signal $X(f)$. The signal waveform we are familiar with is converted into a spectrum drawn by the frequency and Fourier coefficients, as shown in Fig. 2.8. In nature, many fundamental signals are sinusoids, such as the free vibration of an undamped system, the vibration caused by the mass eccentricity of rotating machinery, pure tone signal, alternating current signal, etc. Because of the correspondence between frequency domain analysis and the physical world, people often use frequency domain analysis to decompose measurement signals to help us understand the physical characteristics of the signals. At the same time, through frequency domain analysis, the measurement signal of a complex physical object can be decomposed into a combination of several simple sinusoidal signals, that is, the complex system is decomposed into several simple systems to simplify the problem.

Example 2.1: Spectrum Analysis of Gearbox Vibration Figure 2.9 shows the waveform and frequency spectrum of a gearbox vibration signal. Because it contains multiple frequency components, it is difficult to obtain useful information about the measured object from the signal waveform. However, after frequency domain analysis, the signal is decomposed into a set of sinusoidal signals. From the magnitude spectrum drawn in Fig. 2.9c, it can be clearly seen that the signal is mainly composed of 4 sinusoidal signal components. With further analysis, we can find that

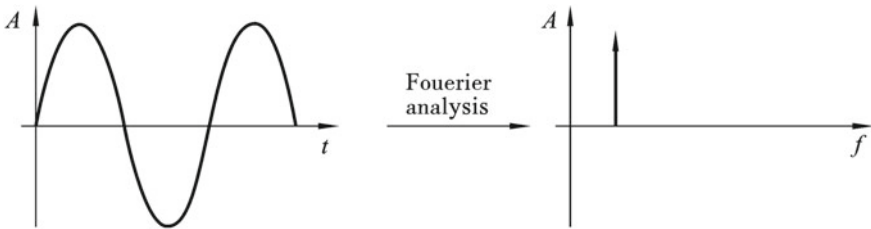


Fig. 2.8 Transformation from time domain into frequency domain

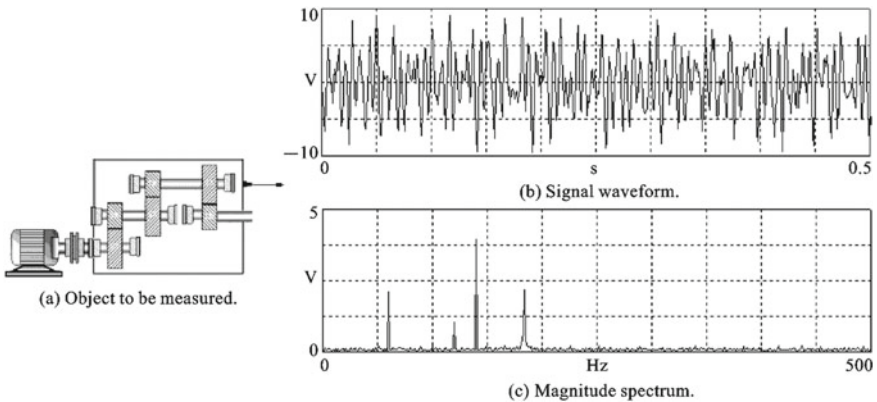


Fig. 2.9 Gearbox vibration signal waveform and frequency spectrum

they respectively correspond to the vibrations of the four rotating shafts. With the help of frequency analysis, we can quickly determine the main vibration source.

2.2 Fourier Series Representation of Periodic Signals

2.2.1 Orthogonal Decomposition of Vectors

Orthogonal decomposition is a method to solve vector calculation problems in physics, and its purpose is to use algebraic operations to solve vector operations. Specifically, any vector in the plane can be decomposed into the sum of two mutually perpendicular vectors. It should be pointed out that the orthogonal vector set is not unique, there can be more than one. You can select a suitable orthogonal vector set to decompose the vector according to your needs, as shown in Fig. 2.10.

From a mathematical point of view, the condition for two vectors in a plane space to be orthogonal (perpendicular to each other) is:

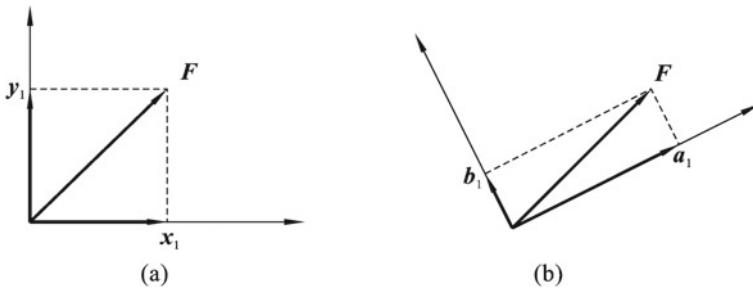


Fig. 2.10 Orthogonal decomposition of vectors

$$\mathbf{X} \cdot \mathbf{Y} = 0 \quad (2.1)$$

In this way, any vector \mathbf{F} in the plane can be decomposed into projections on two orthogonal vectors:

$$\mathbf{F} = x_1 \mathbf{X} + y_1 \mathbf{Y} \quad (2.2)$$

where x_1 and y_1 are decomposition coefficients on \mathbf{X} and \mathbf{Y} .

Similarly, for a vector \mathbf{F} in 3-dimensional space, three orthogonal vectors $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ should be used to form a 3-dimensional orthogonal vector set $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ to decompose it. It is incomplete and insufficient to represent a 3-dimensional vector with a 2-dimensional orthogonal vector set. In the same way, for a vector \mathbf{F} in n -dimensional space, it is necessary to use a orthogonal set with n orthogonal vectors $\mathbf{A}_1, \mathbf{A}_2 \dots \mathbf{A}_n$.

2.2.2 Orthogonal Function

The concept of orthogonal decomposition of vectors can also be extended to signal analysis. Generally speaking, the signal is often expressed as a function of time, so the decomposition of the signal is the decomposition of the time function.

According to the concept of vector orthogonality, we can define the orthogonality of functions. Assuming $f_1(t)$ and $f_2(t)$ are two real functions (signals) defined in the interval of $[t_1, t_2]$, and the condition

$$\int_{t_2}^{t_1} f_1(t) f_2(t) dt = 0 \quad (2.3)$$

is satisfied, then, we say that $f_1(t)$ and $f_2(t)$ are orthogonal in the interval $[t_1, t_2]$. Similarly, expand to the case to multiple functions $f_1(t), f_2(t), \dots, f_n(t)$, they are said

to be mutually orthogonal if:

$$\begin{cases} \int_{t_2}^{t_1} f_i(t) f_j(t) dt = 0, & i \neq j \\ \int_{t_2}^{t_1} f_i(t) f_j(t) dt = K_{ij}, & i = j \end{cases} \quad (2.4)$$

where K_{ij} is a constant. And the function set $\{f_1(t), f_2(t), \dots, f_n(t)\}$ is an orthogonal set in the interval $[t_1, t_2]$. If there is no other non-zero function that is orthogonal to every function in the function set, then the function set is called complete orthogonal function set. Besides, it is called an incomplete orthogonal function set.

Similar to the orthogonal decomposition of vectors, a signal $x(t)$ in the interval $[t_1, t_2]$ can be decomposed by projecting on the function set $\{f_1(t), f_2(t), \dots, f_n(t)\}$:

$$x(t) = c_1 f_1(t) + c_2 f_2(t) + \dots + c_n f_n(t), \quad t_1 \leq t \leq t_2 \quad (2.5)$$

where c_1, c_2, \dots, c_n are decomposition coefficients of $x(t)$.

2.2.3 Orthogonality of Trigonometric Functions

Trigonometric functions were first studied in the eighteenth century in the study of harmonic vibrations. At that time, French mathematician and physicist Joseph Fourier made an in-depth study of trigonometric functions and proposed the trigonometric function set and the famous Fourier transform. Within the interval $[t_0, t_0 + T]$, the trigonometric function set $\{\cos(2\pi i f_0 t), \sin(2\pi j f_0 t)\} (i = 0, 1, 2, \dots, j = 0, 1, 2, \dots)$ satisfies:

$$\begin{aligned} \int_{t_0}^{t_0+T} \cos(2\pi i f_0 t) \cos(2\pi j f_0 t) dt &= \begin{cases} 0 & (i \neq j) \\ \frac{T}{2} & (i = j) \\ T & (i = j = 0) \end{cases} \\ \int_{t_0}^{t_0+T} \sin(2\pi i f_0 t) \sin(2\pi j f_0 t) dt &= \begin{cases} 0 & (i \neq j, \text{ or } i = j = 0) \\ \frac{T}{2} & (i = j) \end{cases} \\ \int_{t_0}^{t_0+T} \sin(2\pi i f_0 t) \cos(2\pi j f_0 t) dt &= 0 \quad i, j \text{ are arbitrary numbers} \end{aligned} \quad (2.6)$$

where $f_0 = 1/T$ is called fundamental frequency. As shown in Eq. (2.6), the trigonometric function set satisfies Eq. (2.4), thus it is an orthogonal function set. It can also

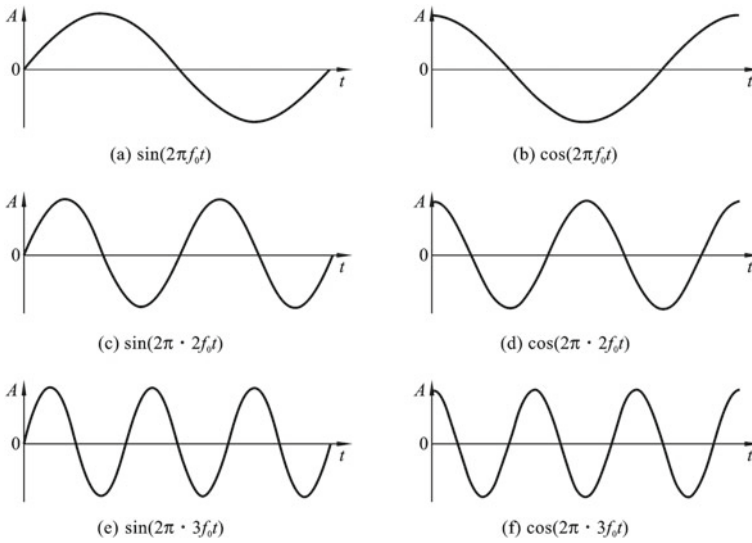


Fig. 2.11 Waveforms of the first three trigonometric functions

be proved that the trigonometric function set is a complete orthogonal function set (students who are interested can try to prove it). Figure 2.11 shows the waveforms of the first three trigonometric functions in the interval $[0, T]$.

Example 2.2: Simulate the Orthogonality of Trigonometric Functions with MATLAB The mathematical equations for verifying the orthogonality of trigonometric functions with frequencies f_0 , $2f_0$, $3f_0$ and $5f_0$, are listed below. The waveforms of the integrands are shown in Fig. 2.12. In the MATLAB codes, the integration is calculated numerically. After running the code, we can see that the integration is almost equal to zero.

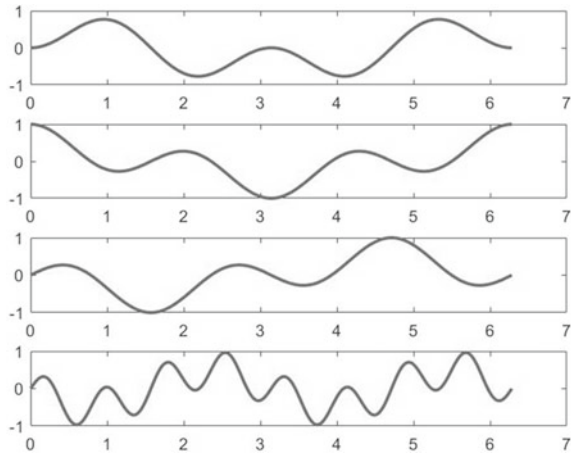
$$\int_0^T \sin(2\pi f_0 t) \cdot \sin(2\pi 2f_0 t) dt = 0$$

$$\int_0^T \cos(2\pi f_0 t) \cdot \cos(2\pi 2f_0 t) dt = 0$$

$$\int_0^T \sin(2\pi f_0 t) \cdot \cos(2\pi 2f_0 t) dt = 0$$

$$\int_0^T \sin(2\pi 3f_0 t) \cdot \cos(2\pi 5f_0 t) dt = 0$$

Fig. 2.12 Waveforms of multiplications of trigonometric functions



MATLAB code for verifying the orthogonality of trigonometric functions:

```

x=linspace(0, 2*pi, 1024);

y1=sin(x);   y2=sin(2*x);   y3=cos(x);

y4=cos(2*x); y5=sin(3*x); y6=cos(5*x);

z1=y1.*y2;   subplot(4,1,1); plot(x,z1,'linewidth',2);

z2=y3.*y4;   subplot(4,1,2); plot(x,z2,'linewidth',2);

z3=y1.*y4;   subplot(4,1,3); plot(x,z3,'linewidth',2);

z4=y5.*y6;   subplot(4,1,4); plot(x,z4,'linewidth',2);

sum(z1*2*pi/1024)

sum(z2*2*pi/1024)

sum(z3*2*pi/1024)

sum(z4*2*pi/1024)

```

2.2.4 Fourier Series in Trigonometric Function Form

Joseph Fourier found that any periodic signal $x(t)$, as long as it satisfies Dirichlet conditions, can be represented by the summation of an infinite series composed of sine functions and cosine functions. The series is called Fourier series. The Dirichlet

conditions are: (1) the function is continuous in any finite interval, or there are only a finite number of discontinuous points of the first kind; (2) in any finite interval, the function can only take a finite maximum or minimum; (3) in any finite interval, the function is integrable.

The Fourier series representation of a periodic signal is:

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} [a_n \cos(2\pi n f_0 t) + b_n \sin(2\pi n f_0 t)] \quad (n = 1, 2, \dots) \quad (2.7)$$

where f_0 is the fundamental frequency, $n f_0$ is the n -th harmonic, a_n and b_n are coefficients of Fourier series, which are usually called Fourier coefficients.

(1) DC component or constant value component a_0

$$a_0 = \frac{2}{T} \int_{-T/2}^{T/2} x(t) dt \quad (2.8)$$

where $T = 1/f_0$ is the period of the signal. If the signal is an odd function, then $a_0 = 0$.

(2) The n -th order cosine coefficients a_n

$$a_n = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \cos(2\pi n f_0 t) dt \quad (n = 1, 2, \dots) \quad (2.9)$$

If the signal is an odd function, then $a_n = 0$. Therefore, the Fourier series coefficients of a periodic odd function only have sine terms.

(3) The n -th order sine coefficients b_n

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \sin(2\pi n f_0 t) dt \quad (n = 1, 2, \dots) \quad (2.10)$$

If the signal is an even function, then $b_n = 0$. Therefore, the Fourier series coefficients of a periodic even function only have cosine terms.

Using the auxiliary angle formula of the trigonometric functions, the sine function and cosine function of the same frequency in Eq. (2.7) can be combined to obtain another form of Fourier series expansion:

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} A_n \cos(2\pi n f_0 t + \varphi_n) \quad (n = 1, 2, \dots) \quad (2.11)$$

where A_n and φ_n are the magnitude and initial phase of each frequency component. They can be calculated:

$$\begin{aligned} A_n &= \sqrt{a_n^2 + b_n^2} \\ \varphi_n &= \arctan \frac{b_n}{a_n} \end{aligned} \quad (2.12)$$

2.2.5 Spectrum of Periodic Signals

In engineering, it is customary to express the Fourier series decomposition results graphically. There are several ways to draw the graph as shown in Fig. 2.13. Take each harmonic frequency f_n as the abscissa and Fourier coefficients a_n and b_n as the ordinate, the graph is called the real and imaginary spectra of the signal; take each harmonic frequency f_n as the abscissa and A_n and φ_n as the ordinate, the graph is called magnitude spectrum and phase spectrum; take each harmonic frequency f_n as the abscissa and A_n^2 as the ordinate, the graph is called power spectrum. Spectra of the signal reflect the composition of the various frequency components, and are commonly used signal analysis.

Example 2.3: Draw the Real, Imaginary, Magnitude, Phase Spectrum and Power Spectra of the Following Signal

$$x(t) = 4 \sin(10\pi t) + 2 \cos(20\pi t)$$

As we can see from Fig. 2.14, $x(t)$ is a periodic signal. Its period can be inferred from the mathematical expression, which is $T = 0.2$. The corresponding fundamental frequency is $f_0 = 5$ Hz. According to the Fourier series decomposition equation, the non-zero coefficients are calculated as

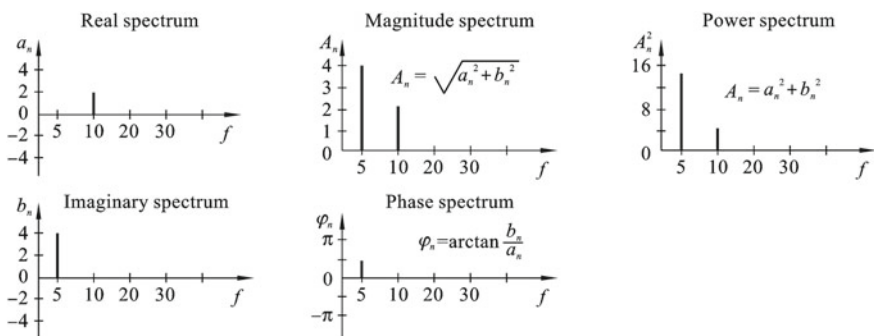
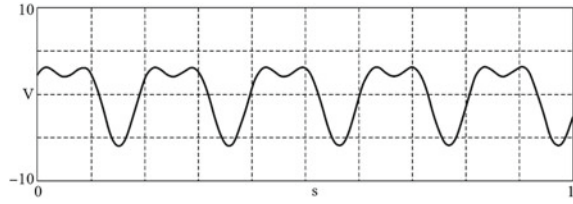


Fig. 2.13 Spectra of periodic signal

Fig. 2.14 Signal waveform

$$b_1 = \frac{2}{0.2} \int_{-0.1}^{0.1} [4 \sin(10\pi t) + 2 \cos(20\pi t)] \cdot \sin(2\pi \cdot 5 \cdot t) dt = 4$$

$$a_2 = \frac{2}{0.2} \int_{-0.1}^{0.1} [4 \sin(10\pi t) + 2 \cos(20\pi t)] \cdot \cos(2\pi \cdot 10 \cdot t) dt = 2$$

According to Eq. (2.12), we get:

$$A_1 = 4, \varphi_1 = \pi/2$$

$$A_2 = 2, \varphi_2 = 0$$

The real spectrum can be drawn from the calculated coefficient a_n , and the imaginary spectrum can be drawn from the coefficient b_n , as shown in Fig. 2.15.

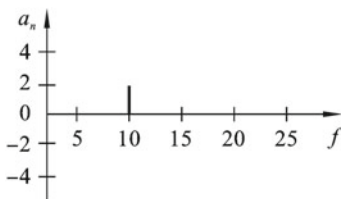
As shown in Fig. 2.16, the magnitude spectrum and phase spectrum can be drawn according to A_n and φ_n respectively.

The power spectrum can be drawn according to A_n^2 , as shown in Fig. 2.17.

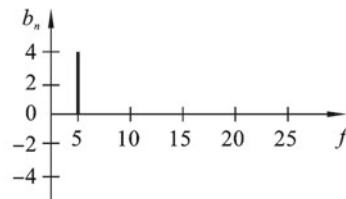
Example 2.4: Find the Fourier Series of the Square Wave Signal in the Following Figure and Draw Its Magnitude Spectrum The mathematical expression of the square wave signal in one period is (Fig. 2.18)

$$x(t) = \begin{cases} A, & 0 \leq t \leq T/2 \\ -A, & -T/2 \leq t < 0 \end{cases}$$

According to Eqs. (2.8)–(2.10), we get



(a) Real spectrum.



(b) Imaginary spectrum.

Fig. 2.15 Real and imaginary spectra of the signal

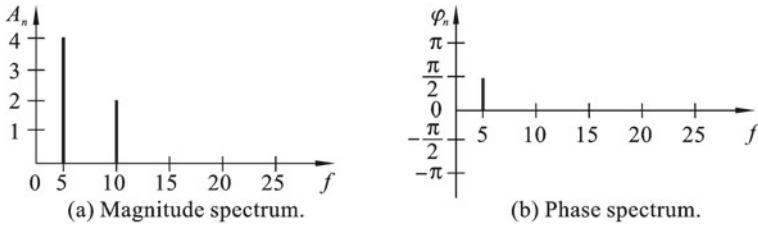


Fig. 2.16 Magnitude and phase spectra of the signal

Fig. 2.17 Power spectrum of the signal

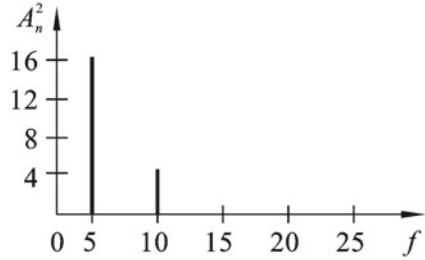
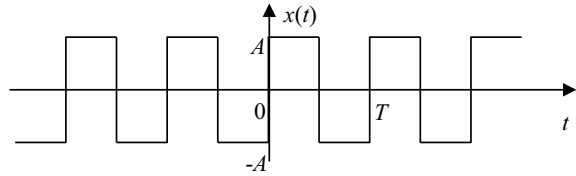


Fig. 2.18 Waveform of the square wave signal

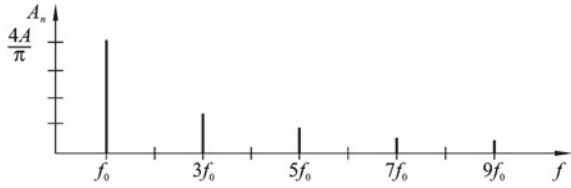


$$a_0 = \frac{2}{T} \int_{-T/2}^{T/2} x(t) dt = \frac{2}{T} \int_0^{T/2} A dt - \frac{2}{T} \int_{-T/2}^0 A dt = 0$$

$$\begin{aligned} a_n &= \frac{2}{T} \int_{-T/2}^{T/2} x(t) \cos(2\pi n f_0 t) dt \\ &= \frac{2}{T} \int_0^{T/2} A \cos(2\pi n f_0 t) dt + \frac{2}{T} \int_{-T/2}^0 -A \cos(2\pi n f_0 t) dt = 0 \end{aligned}$$

$$b_n = \frac{2}{T} \int_{-T/2}^{T/2} x(t) \sin(2\pi n f_0 t) dt$$

Fig. 2.19 Magnitude spectrum of the square wave signal



$$\begin{aligned}
 &= \frac{2}{T} \int_0^{T/2} A \sin(2\pi n f_0 t) dt + \frac{2}{T} \int_{-T/2}^0 -A \sin(2\pi n f_0 t) dt \\
 &= \frac{2A}{n\pi} [1 - \cos(n\pi)] \\
 &= \begin{cases} \frac{4A}{n\pi}, n = 1, 3, 5, \dots \\ 0, n = 2, 4, 6, \dots \end{cases}
 \end{aligned}$$

where $f_0 = 1/T$ is the fundamental frequency. Substituting the coefficients into Eq. (2.7), the Fourier series of the square wave signal can be obtained:

$$x(t) = \frac{4A}{\pi} \left[\sin(2\pi f_0 t) + \frac{1}{3} \sin(2\pi 3 f_0 t) + \frac{1}{5} \sin(2\pi 5 f_0 t) + \frac{1}{7} \sin(2\pi 7 f_0 t) + \dots \right]$$

Calculating A_n according to Eq. (2.12), its magnitude spectrum can be drawn in Fig. 2.19.

It can be seen from Examples 2.1 and 2.2 that the frequency spectrum of periodic signals has the following characteristics: first, the frequency spectrum of the continuous periodic signal is a discrete spectrum; second, the spectral line only appears at the fundamental frequency and its harmonics.

2.2.6 Synthesis of Periodic Signals

Any periodic signal can be decomposed into the sum of harmonic sine and cosine signals. Therefore, periodic signals can be synthesized with sine and cosine signals.

Square wave signal: $\sin(x) + \sin(3x)/3 + \sin(5x)/5 + \dots$

Sawtooth signal: $\sin(x) + \sin(2x)/2 + \sin(3x)/3 + \dots$

Triangular signal: $\sin(x) - \sin(3x)/9 + \sin(5x)/25 - \sin(7x)/49 \dots$

Example 2.5: Synthesize a Square Wave Signal in MATLAB The following MATLAB code is used to synthesize a square wave with the summation of 15 sinusoids. With the “pause” function in the code, we can see the dynamic change of the waveform as more and more sinusoids are added (Fig. 2.20).

```

x=linspace(0, 6*pi, 1000);

z=zeros(1,1000); k=1;

for i=1:15

    y=(1/k)*sin(k*x); z=z+y;

    plot(x,z,'linewidth',1)

    xlim([0,6*3.14]); ylim([-1.2,1.2]);

    grid('on'); pause(1); k=k+2;

end

```

Example 2.6: Synthesizing a Triangular Wave Signal in MATLAB The following MATLAB code is used to synthesize a triangular wave with 4 sinusoids (Fig. 2.21).

Fig. 2.20 Generated square wave signal by synthesis

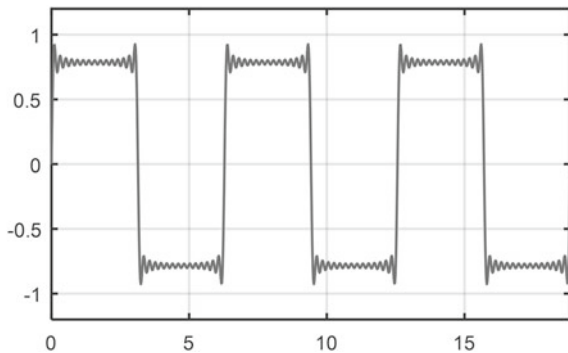
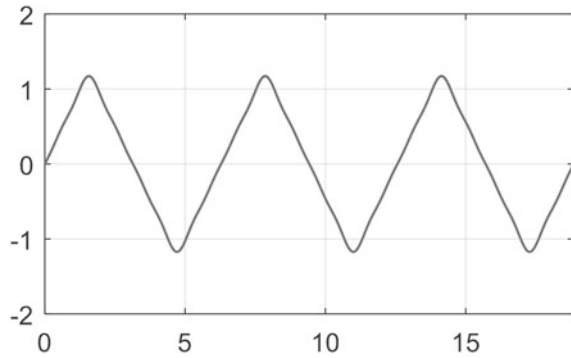


Fig. 2.21 Generated triangular wave signal by synthesis



```

x=linspace(0, 6*pi, 1000);

y1=sin(x);      y2=sin(3*x)/9;

y3=sin(5*x)/25;  y4=sin(7*x)/49;

y=y1-y2+y3-y4;

plot(x,y,'linewidth',1);

xlim([0,6*3.14]); grid on

```

2.2.7 Walsh Orthogonal Function Set

The communication theory was based on the sine and cosine function system in the past. With the progress in semiconductor electronics, modern digital communication technology with 0 and 1 binaries as carrier has gradually developed. The Walsh function is a function set introduced by the American mathematician J. L. Walsh in 1923. Its value can only take $+1$ and -1 . It is a complete set of orthogonal functions in the interval $[0, 1]$. The image itself can be seen as a waveform of a binary signal. The simplicity of the values is particularly suitable for processing digital signals, so it exhibits great application in modern communications. Figure 2.22 shows the Walsh functions of the first four orders.

Besides, there are many other orthogonal function sets, such as Legendre function set, Bessel function set, Wavelet function set, etc., all of which can be used to decompose signals.

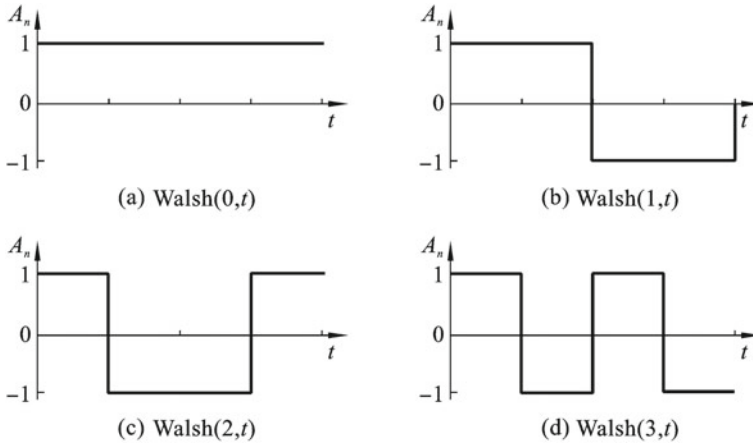


Fig. 2.22 Walsh function waveforms

2.2.8 Fourier Series in Complex Form

The advantage of the Fourier series in the form of trigonometric functions is that the physical meaning is clear. It can reveal the harmonic components contained in the signal very well. But when the signal $x(t)$ is more complex, the integral operation for calculating the Fourier coefficients is very difficult to implement. To facilitate the mathematical operations, Fourier series in complex form can be used.

According to Euler's formula, we have:

$$\begin{aligned}\cos(2\pi ft) &= \frac{1}{2}(e^{-j2\pi ft} + e^{j2\pi ft}) \\ \sin(2\pi ft) &= \frac{j}{2}(e^{-j2\pi ft} - e^{j2\pi ft})\end{aligned}\quad (2.13)$$

Substituting them into Eq. (2.7), we get:

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[\frac{1}{2}(a_n - jb_n)e^{jn2\pi f_0 t} + \frac{1}{2}(a_n + jb_n)e^{-jn2\pi f_0 t} \right] \quad (2.14)$$

Assuming:

$$\begin{aligned}C_0 &= \frac{1}{2}a_0 \\ C_n &= \frac{1}{2}(a_n - jb_n) \\ C_{-n} &= \frac{1}{2}(a_n + jb_n)\end{aligned}\quad (2.15)$$

then we have:

$$x(t) = C_0 + \sum_{n=1}^{\infty} [C_n e^{jn2\pi f_0 t} + C_{-n} e^{-jn2\pi f_0 t}] = \sum_{n=-\infty}^{\infty} C_n e^{jn2\pi f_0 t} \quad (2.16)$$

This is the Fourier series expansion in complex exponential form, where C_n represents the complex amplitude of the signal, which is:

$$C_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-jn2\pi f_0 t} dt \quad (n = 0, \pm 1, \pm 2, \dots) \quad (2.17)$$

Equations (2.16) and (2.17) are referred as the synthesis equation and analysis equation respectively. Usually, C_n is a complex number, which can be denoted as:

$$C_n = |C_n| e^{j\varphi_n} = \text{Re}(C_n) + j\text{Im}(C_n) \quad (2.18)$$

where $|C_n|$ and φ_n are magnitude and phase of the complex number C_n ; $\text{Re}(C_n)$ and $\text{Im}(C_n)$ are real and imaginary components of C_n .

According to the Fourier coefficients in complex form C_n , the real and imaginary spectra, magnitude and phase spectra and power spectrum of the signal can also be drawn. One thing to be noted is that, in the form of trigonometric functions, the subscripts of a_n and b_n are always positive and the corresponding frequency range is $(0, \infty)$, while in the complex form, the subscripts of C_n can be negative, and the corresponding frequency range is expanded to $(-\infty, \infty)$. Therefore, the frequency spectrum of the Fourier coefficients in the trigonometric function form is a one-sided spectrum, and the frequency spectrum of the Fourier coefficients in the complex form is a two-sided spectrum. Moreover, C_n has conjugate symmetry about n , and the magnitude of a_n ($n \neq 0$) equals the sum of C_n at positive and negative n . If the signal is an odd function, then $C_n = -C_{-n} = -\frac{1}{2}jb_n$ (C_n is a pure imaginary number), $\varphi_n = -\frac{\pi}{2}$; if the signal is an even function, then $C_n = \frac{a_n - jb_n}{2} = \frac{a_n}{2} = C_{-n}$ (C_n is a real number), $\varphi_n = 0$.

For example, for a simple cosine signal

$$x(t) = A \cos(2\pi f_0 t)$$

there is only one Fourier coefficient $a_1 = A$ in the form of trigonometric functions. And in the complex form, it is decomposed as:

$$x(t) = \frac{A}{2} (e^{-j2\pi f_0 t} + e^{j2\pi f_0 t})$$

where the Fourier coefficients become $C_{-1} = A/2$ and $C_1 = A/2$. Figure 2.23 shows the magnitude spectra in trigonometric form and complex exponential form.

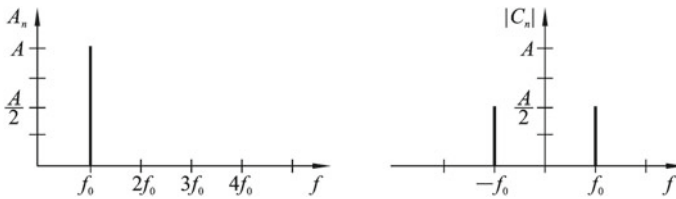


Fig. 2.23 Magnitude spectra in trigonometric form and complex form

It can be seen that the spectrum drawn by the complex Fourier coefficients is a two-sided spectrum, and negative frequencies appear in the figure. The magnitude of the two-sided spectrum is half of the one-sided spectrum. Due to the lack of physical meaning of negative frequencies, two-sided spectra are mainly used in mathematical proofs and derivations. Usually, in engineering practice, they are converted to one-sided spectra by adding the magnitudes of positive and negative frequencies.

A summary of the Fourier spectrum of periodic signals and its properties are listed below:

- (1) $\{C_n\}$ are the coefficients of complex Fourier series, also referred as complex Fourier coefficients and Fourier series spectrum.
- (2) $\{|C_n|\}$ is the magnitude spectrum of complex Fourier series.
- (3) $\{\varphi_n\}$ is the phase spectrum of complex Fourier series.
- (4) The Fourier series spectrum of a periodic signal only has values at the angular frequency $n\omega_0$ (or frequency nf_0). The spectrum will never appear in between the harmonics.
- (5) Fourier series spectrum is also called discrete Fourier spectrum, the discrete interval is $f_0 = 1/T_0$.
- (6) The discrete lines in the magnitude spectrum and phase spectrum that represent the corresponding values are called spectrum lines.
- (7) The line connecting the peaks of the spectrum lines is called envelope, which reflects the variation of the Fourier series spectrum, magnitude spectrum and phase spectrum with frequency.
- (8) a_n is a single-sided spectrum, which represents the actual component magnitude of the signal at each frequency.
- (9) C_n is a two-sided spectrum. Its negative frequency term does not exist in practice. The actual magnitude is the sum of the values at positive and negative frequencies.

Figure 2.24 shows the periodic rectangular pulse sequence and its Fourier series spectrum. The characteristics of the Fourier series spectrum are:

- (1) The envelope of the spectrum is a sinc function; the zero crossing points of the envelope of the spectrum are $f = nf_0 = k/\tau$, where τ is the pulse width.
- (2) In the frequency domain, the energy is concentrated within the first zero crossing.

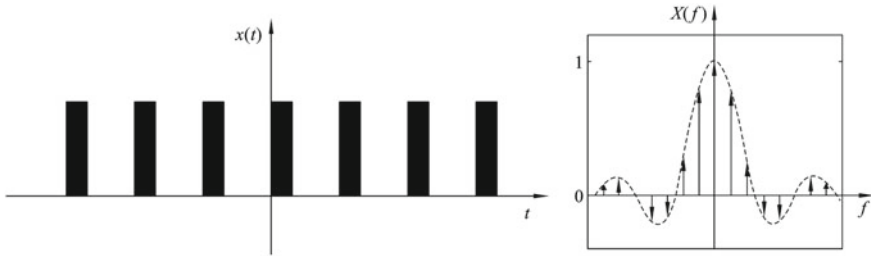


Fig. 2.24 Periodic rectangular pulse sequence and its Fourier series spectrum

- (3) Bandwidth is $\beta_f = 1/\tau$ in frequency form or $\beta_0 = 2\pi/\tau$ in angular frequency form. It is only related to the pulse width τ , and irrelevant to pulse height and period of the signal.

2.2.9 Gibbs Phenomenon

It can be seen from the above introduction that a periodic signal can be decomposed into a series of sine and cosine signals with a Fourier series; conversely, a series of sine and cosine signals can be synthesized to obtain the corresponding periodic signal. In order to further understand this conversion relationship between time domain and frequency domain, it is necessary to explain the so-called Gibbs phenomenon.

If a periodic function with jump discontinuities (such as a square wave signal) is synthesized by finite terms of Fourier coefficients, there will be large oscillations near the jump. This phenomenon is called Gibbs phenomenon, which was discovered by Henry Wilbraham in 1848 and rediscovered by J. Willard Gibbs in 1899. The more the number of terms is selected, the smaller the oscillation. The Gibbs phenomenon is caused by the inability of the expansion to converge uniformly in the neighborhood of the discontinuity. Thus, this phenomenon will not die out even if the number of terms goes to infinity. It will approach a finite limit of about 9% of the total jump value.

Example 2.7: Synthesize a Square Wave Signal with Fourier Series The results of synthesizing a square wave signal with the first 1, 3 and 5 Fourier coefficients are shown in Fig. 2.25. We can clearly see the Gibbs phenomenon in the synthesized signal, i.e. the oscillation at discontinuities. The larger the number of Fourier coefficients used in the synthesis, the closer the waveform to the original signal.

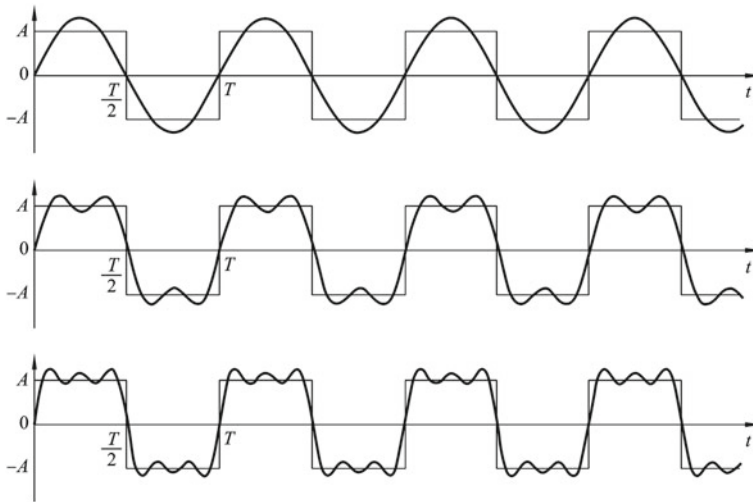


Fig. 2.25 Synthesis of a square wave signal with different number of Fourier coefficients

2.2.10 Parseval's Theorem

Parseval's theorem shows the relationship between the signal energy in time domain and the signal energy in frequency domain. It indicates the power or energy in time domain is equivalent to that in frequency domain. For two-sided spectrum, the Parseval's theorem is defined as:

$$\frac{1}{T} \int_0^T x^2(t) dt = \sum_{n=-\infty}^{\infty} |C_n|^2 \quad (2.19)$$

For one-sided spectrum, the Parseval's theorem is defined as:

$$\frac{1}{T} \int_0^T x^2(t) dt = \sum_{n=0}^{\infty} A_n^2/2 \quad (2.20)$$

The Parseval's theorem can be proved by substituting the equation of complex Fourier series into Eq. (2.19), and using the orthogonal relationship between the complex exponential functions.

For a signal $x(t) = A \cos(2\pi f_0 t) + B \cos(4\pi f_0 t)$, the power can be calculated according to the left-hand side of Eq. (2.19):

$$\begin{aligned}
P &= \frac{1}{T} \int_0^T x^2(t) dt \\
&= \frac{1}{T} \int_0^T [A \cos(2\pi f_0 t) + B \cos(4\pi f_0 t)] \cdot [A \cos(2\pi f_0 t) + B \cos(4\pi f_0 t)] dt
\end{aligned}$$

According to the orthogonality of trigonometric functions, we get:

$$\begin{aligned}
P &= \frac{1}{T} \int_0^T [A \cos(2\pi f_0 t)]^2 dt + \frac{1}{T} \int_0^T [B \cos(4\pi f_0 t)]^2 dt \\
&= A^2/2 + B^2/2
\end{aligned}$$

The complex Fourier coefficients are $C_{-1} = A/2$, $C_1 = A/2$, $C_{-2} = B/2$, $C_2 = B/2$. The power in frequency domain can be calculated according to the right-hand side of Eq. (2.19):

$$\begin{aligned}
P &= C_{-2}^2 + C_{-1}^2 + C_1^2 + C_2^2 \\
&= A^2/2 + B^2/2
\end{aligned}$$

As we can see, the powers in time domain and frequency domain are the same.

2.3 Fourier Transform of Aperiodic Signals

2.3.1 Fourier Integral

Fourier integral is developed on the basis of Fourier series, and is mainly used for spectrum analysis of aperiodic signals. The Fourier integral can be understood as follows: an aperiodic signal can be regarded as a periodic signal with its period goes to infinity $T \rightarrow \infty$; at the same time, the fundamental frequency goes to an infinitesimal $f_0 \rightarrow df$, thus the spectrum contains all frequency components from zero to infinity; correspondingly, the sum of Fourier series becomes the Fourier integral $\sum \rightarrow \int$. A brief derivation is shown below.

The Fourier expansion of the periodic signal $x(t)$ in complex exponential form is:

$$x(t) = \sum_{n=-\infty}^{\infty} C_n e^{jn2\pi f_0 t} \quad (2.21)$$

The Fourier coefficients are:

$$C_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-jn2\pi f_0 t} dt \quad (2.22)$$

Substitute C_n into the Fourier expansion, we get:

$$\begin{aligned} x(t) &= \sum_{n=-\infty}^{\infty} \left[\frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-jn2\pi f_0 t} dt \right] e^{jn2\pi f_0 t} \\ &= \sum_{n=-\infty}^{\infty} \left[\int_{-T/2}^{T/2} x(t) e^{-jn2\pi f_0 t} dt \right] e^{jn2\pi f_0 t} \frac{1}{T} \end{aligned} \quad (2.23)$$

If $T \rightarrow \infty$, then $\int_{-T/2}^{T/2} \rightarrow \int_{-\infty}^{\infty}$, $f_0 = \frac{1}{T} \rightarrow df$, $nf_0 \rightarrow f$ and $\sum \rightarrow \int$. Substituting them into Eq. (2.23), we get:

$$x(t) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x(t) e^{-j2\pi f t} dt \right] e^{j2\pi f t} df \quad (2.24)$$

Notate the part in square bracket as:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi f t} dt \quad (2.25)$$

Equation (2.24) can be written as:

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi f t} df \quad (2.26)$$

Equation (2.25) is the equation for Fourier transform (FT) and Eq. (2.26) is the equation for inverse Fourier transform (IFT). They form a Fourier transform pair. Equations (2.25) and (2.26) are referred as the analysis equation and synthesis equation of Fourier transform. Sometimes, the Fourier transform is also expressed in angular frequency:

$$X(j\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt$$

Properties of FT and IFT are as follows:

- (1) $e^{-j2\pi ft}$ (or $e^{-j\omega t}$) is called the kernel function of FT, and $e^{j2\pi ft}$ (or $e^{j\omega t}$) is the kernel function of IFT.
- (2) FT and IFT are unique. If the FT of two functions are equal, then the two functions must be equal.
- (3) FT is reversible. If $F[x(t)] = X(f)$, then $F^{-1}[X(f)] = x(t)$, and vice versa.
- (4) The Fourier transform of a signal is generally a complex-valued function, which can be written as $X(j\omega) = |X(j\omega)|e^{j\varphi(j\omega)}$. $|X(j\omega)|$ is the magnitude spectral density function, abbreviated as the magnitude spectrum, which represents the magnitude-frequency characteristic of the signal. $\varphi(j\omega)$ is called the phase spectral density function, abbreviated as the phase spectrum, which represents the phase-frequency characteristic of the signal.
- (5) The FT spectrum can be decomposed into real and imaginary parts: $X(j\omega) = X_R(j\omega) + jX_I(j\omega)$, and

$$|X(j\omega)| = \sqrt{X_R^2(j\omega) + X_I^2(j\omega)}, \quad \varphi(j\omega) = \arctan \frac{X_R(j\omega)}{X_I(j\omega)} \quad (2.27)$$

$$X_R(j\omega) = |X(j\omega)| \cos(\varphi(j\omega)), \quad X_I(j\omega) = |X(j\omega)| \sin(\varphi(j\omega)) \quad (2.28)$$

- (6) FT and IFT are defined as:

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft} dt, \quad x(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft} df \quad (2.29)$$

2.3.2 Spectrum of Aperiodic Signal

Similar to the frequency analysis of periodic signals, the Fourier integration results of aperiodic signals can also be graphically represented, i.e. real and imaginary spectra, magnitude and phase spectra and power spectrum. The difference is that, since the period of the aperiodic signal goes to infinity $T \rightarrow \infty$, the fundamental frequency goes to an infinitesimal $f_0 \rightarrow df$. The harmonic frequency components are:

$$f = n \cdot df \quad (2.30)$$

Since df is an infinitesimal quantity, f becomes a continuous variable. The spectrum contains all frequency components from negative infinity to positive infinity, so the spectrum of aperiodic signals is continuous.

In addition, from the inverse Fourier transform equation listed in Eq. (2.26), it can be inferred that the magnitude of each frequency component is $X(f)df$. Since df is an infinitesimal quantity, $X(f)df$ is also an infinitesimal quantity. In this situation, the frequency spectrum can no longer be expressed in magnitude, but must be expressed in magnitude density per unit frequency.

$$X(f)df/df = X(f) \quad (2.31)$$

Therefore, the aperiodic signal spectrum represents the magnitude density. In fact, it should be correctly called the spectral density. But, customarily, we still use the phrase “magnitude spectrum” to refer it. From Eq. (2.31), we can see that the spectral density of aperiodic signals is the coefficient of Fourier integral, which is a complex number and can be expressed as:

$$X(f) = X_R(f) + jX_I(f) = |X(f)|e^{j\varphi(f)} \quad (2.32)$$

where $X_R(f)$ and $X_I(f)$ are the real and imaginary components of $X(f)$, $|X(f)|$ is the magnitude of $X(f)$, and $\varphi(f)$ is the phase of $X(f)$:

$$\begin{aligned} |X(f)| &= \sqrt{X_R^2(f) + X_I^2(f)} \\ \varphi(f) &= \arctan\left(\frac{X_I(f)}{X_R(f)}\right) \end{aligned} \quad (2.33)$$

The real and imaginary spectra can be drawn with $X_R(f)$ and $X_I(f)$, the magnitude and phase spectra can be drawn with $|X(f)|$ and $\varphi(f)$, and the power spectrum can be drawn with $|X(f)|^2$. The frequency range of Fourier integral $X(f)$ is $(-\infty, \infty)$, thus the spectrum is a two-sided spectrum. In engineering applications, generally only the positive frequency of the spectrum is considered, and the frequency range is $[0, \infty)$. This kind of frequency spectrum is called one-sided spectrum, and its relationship with two-sided spectrum is:

$$X_{\text{one}}(f) = 2X_{\text{two}}(f) \quad (2.34)$$

2.4 Fourier Transform of Typical Signals

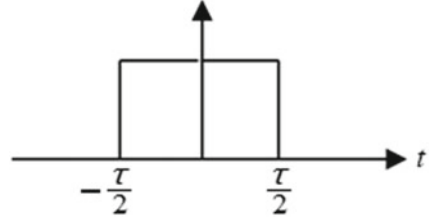
The Fourier transforms of some typical signals, including rectangular pulse signal, unit impulse signal, sign signal, unit step function, periodic impulse train, rectangular window function and sinusoids, are analyzed in this section.

1. Rectangular pulse signal

Example 2.8: Find the Spectra of the Rectangular Pulse Signal Shown in Fig. 2.26 The rectangular pulse signal is mathematically expressed as:

$$u(t) = \begin{cases} 1, & |t| \leq T/2 \\ 0, & |t| > T/2 \end{cases}$$

Fig. 2.26 Waveform of rectangular pulse signal



Substituting it into the equation of Fourier transform, we obtain:

$$\begin{aligned}
 X(f) &= \int_{-\infty}^{\infty} u(t)e^{-j2\pi ft} dt = \int_{-T/2}^{T/2} e^{-j2\pi ft} dt \\
 &= \frac{1}{-j2\pi f} (e^{-j2\pi fT/2} - e^{j2\pi fT/2}) \\
 &= \frac{\sin(\pi fT)}{\pi f} \\
 &= T \frac{\sin(\pi fT)}{\pi fT}
 \end{aligned}$$

There is a special name for $\sin(x)/x$ in mathematics, i.e. sinc(x) function. From the above formula, we get:

$$\begin{aligned}
 X_R(f) &= T \frac{\sin(\pi fT)}{\pi fT} \\
 X_I(f) &= 0 \\
 |X(f)| &= \sqrt{\left[\frac{\sin(\pi fT)}{\pi fT} \right]^2} \\
 \varphi(f) &= \begin{cases} 0, & X_R(f) > 0 \\ \pm\pi, & X_R(f) \leq 0 \end{cases}
 \end{aligned}$$

The real and imaginary spectra can be drawn with $X_R(f)$ and $X_I(f)$, as shown in Fig. 2.27.

The magnitude and phase spectra can be drawn with $|X(f)|$ and $\varphi(f)$, as shown in Fig. 2.28.

The power spectrum can be drawn with $|X(f)|^2$, as shown in Fig. 2.29.

The Fourier transform of rectangular pulse signal has the following properties:

- (1) Its Fourier transform is sinc function, the function value at the origin is equal to the area of the rectangular pulse;

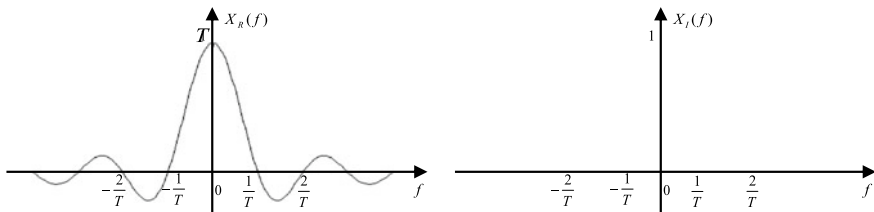


Fig. 2.27 Real and imaginary spectra of the rectangular pulse signal

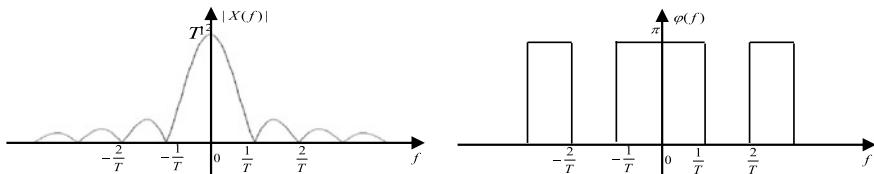
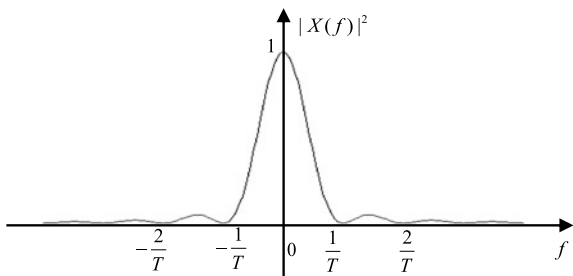


Fig. 2.28 Magnitude and phase spectra of the rectangular pulse signal

Fig. 2.29 Power spectrum of the rectangular pulse signal



- (2) The zero crossing points of the Fourier transform result are $f = k/T$ ($k \neq 0$), where T is the pulse width;
- (3) The energy in the frequency domain is concentrated within the first zero-crossing interval $f \in [-1/T, 1/T]$;
- (4) Bandwidth is $B_f = 1/T$ in frequency form or $B_0 = 2\pi/T$ in angular frequency form. It is only related to the pulse width T , and irrelevant to pulse height and period of the signal. As shown in Fig. 2.30, the equivalent pulse width and bandwidth of the signal are $T = F(0)/f(0)$ and $B_f = 1/T$ respectively.

2. δ function

δ function is also called unit impulse function. Its detailed definition can be referred in Chap. 1. The Fourier transform of δ function is:

$$X(f) = \int_{-\infty}^{\infty} \delta(t) e^{-j2\pi f t} dt = e^0 = 1 \quad (2.35)$$

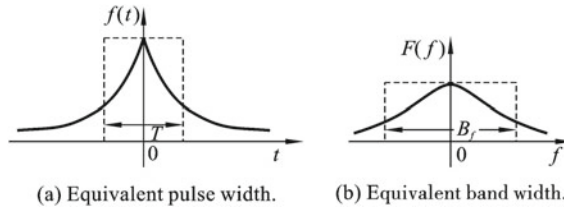


Fig. 2.30 Equivalent pulse width and bandwidth of the signal

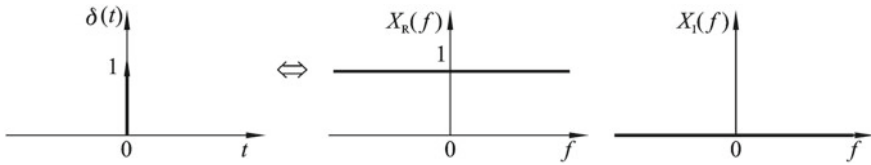


Fig. 2.31 Waveform and corresponding real and imaginary spectra of δ function

Its waveform and corresponding real and imaginary spectra are shown in Fig. 2.31.

3. Sign function

Sign function is mathematically defined as:

$$\text{sgn}(t) = \begin{cases} 1, & t > 0 \\ 0, & t = 0 \\ -1, & t < 0 \end{cases} \quad (2.36)$$

The sign function does not satisfy the absolutely integrable condition, but there is Fourier transform for it. Differentiate the sign function, we get:

$$\frac{d}{dt} \text{sgn}(t) = 2\delta(t)$$

According to the differential property of the Fourier transform introduced in Sect. 2.5, the Fourier transform of the sign function is

$$X(f) = \frac{2}{j2\pi f} = -j \frac{1}{\pi f} \quad (2.37)$$

Figure 2.32 shows the waveform and corresponding real and imaginary spectra of sign function.

4. Unit step function

Unit step function is mathematically defined as:

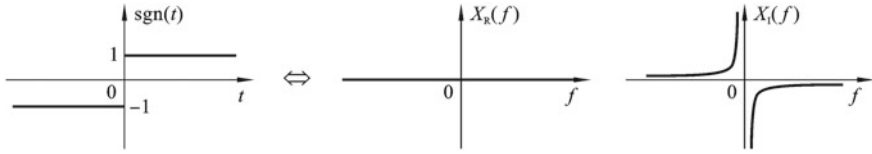


Fig. 2.32 Waveform and corresponding real and imaginary spectra of sign function

$$u(t) = \begin{cases} 0, & t < 0 \\ 1, & t \geq 0 \end{cases} \quad (2.38)$$

It can be re-written in the form of sign function:

$$u(t) = \frac{1}{2} + \frac{1}{2}\text{sgn}(t) \quad (2.39)$$

According to the linear superposition property of the Fourier transform, we get:

$$F[u(t)] = F\left[\frac{1}{2}\right] + F\left[\frac{1}{2}\text{sgn}(t)\right] \quad (2.40)$$

It can be known from the symmetry properties of the Fourier transform and the properties of the δ function, the Fourier transform of the constant k is $k\delta(f)$. Then, we get:

$$F[u(t)] = \frac{1}{2}\delta(f) - j\frac{1}{2\pi f} \quad (2.41)$$

Figure 2.33 shows the waveform and corresponding real and imaginary spectra of unit step function.

5. Periodic impulse train

Periodic impulse train is mathematically defined as:

$$x(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT) \quad (2.42)$$

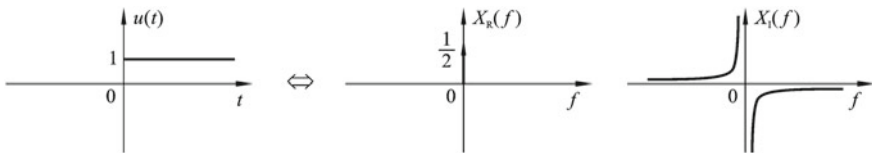


Fig. 2.33 Waveform and corresponding real and imaginary spectra of unit step function

where T is the period. It can be expanded with Fourier series:

$$x(t) = \sum_{n=-\infty}^{\infty} C_n e^{j2\pi n f_0 t} \quad (2.43)$$

where $f_0 = 1/T$ is the fundamental frequency, C_n are the Fourier coefficients:

$$C_n = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-j2\pi n f_0 t} dt = \frac{1}{T} \int_{-T/2}^{T/2} \delta(t) e^{-j2\pi n f_0 t} dt = \frac{1}{T} \quad (2.44)$$

Take the Fourier transform of Eq. (2.43), we get:

$$X(f) = F \left[\sum_{n=-\infty}^{\infty} \frac{1}{T} e^{j2\pi n f_0 t} \right] \quad (2.45)$$

According to the frequency shifting property of the Fourier transform, we get:

$$X(f) = \frac{1}{T} \left[\sum_{n=-\infty}^{\infty} \delta(f - n f_0) \right] \quad (2.46)$$

or

$$X(j\omega) = 2\pi \sum_{n=-\infty}^{\infty} \frac{1}{T} \delta(\omega - n\omega_0) = \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0)$$

where ω is angular frequency. Figure 2.34 shows the waveform and corresponding real and imaginary spectra of periodic impulse train.

6. Rectangular window function

The rectangular window function is a single rectangular pulse, and it is mathematically definite as:

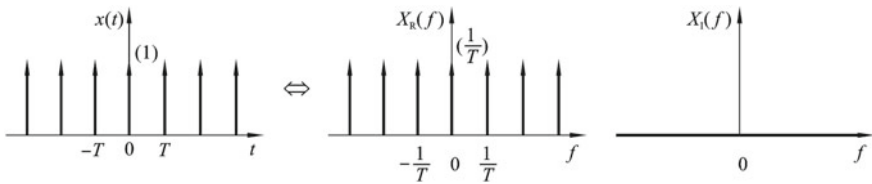


Fig. 2.34 Waveform and corresponding real and imaginary spectra of periodic impulse train

$$u(t) = \begin{cases} 1, & |t| \leq T/2 \\ 0, & |t| > T/2 \end{cases} \quad (2.47)$$

In Example 2.8, we have discussed a single rectangular pulse function, substituting it into the Fourier transform equation, the Fourier transform result can be obtained:

$$X(f) = \frac{\sin(\pi f T)}{\pi f} \quad (2.48)$$

Figure 2.35 shows the waveform and corresponding real and imaginary spectra of rectangular window function.

7. Sine and cosine signals

Sine and cosine functions are defined as:

$$x(t) = \sin(2\pi f_0 t) \quad (2.49)$$

$$y(t) = \cos(2\pi f_0 t) \quad (2.50)$$

where f_0 is the frequency. According to Euler's formula, sine and cosine functions can be written as:

$$x(t) = j\frac{1}{2}(e^{-j2\pi f_0 t} - e^{j2\pi f_0 t}) \quad (2.51)$$

$$y(t) = \frac{1}{2}(e^{-j2\pi f_0 t} + e^{j2\pi f_0 t}) \quad (2.52)$$

According to the frequency shifting property of Fourier transform, we get:

$$X(f) = j\frac{1}{2}[\delta(f + f_0) - \delta(f - f_0)] \quad (2.53)$$

$$Y(f) = \frac{1}{2}[\delta(f + f_0) + \delta(f - f_0)] \quad (2.54)$$

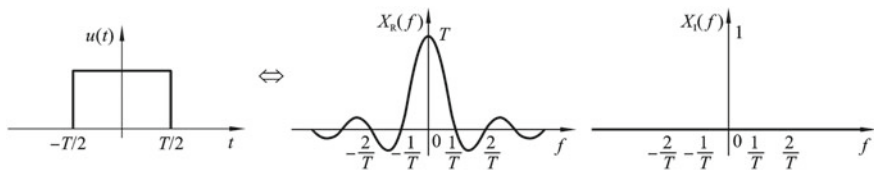


Fig. 2.35 Waveform and corresponding real and imaginary spectra of rectangular window function

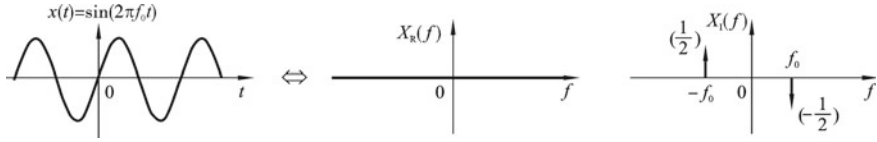


Fig. 2.36 Waveform and corresponding real and imaginary spectra of sine function

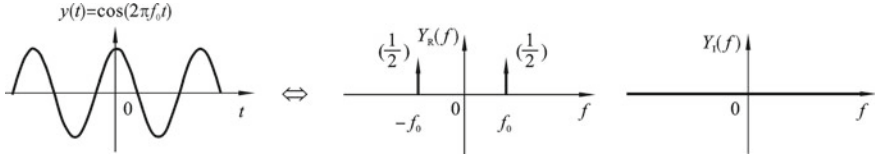


Fig. 2.37 Waveform and corresponding real and imaginary spectra of cosine function

Figures 2.36 and 2.37 are waveforms and corresponding real and imaginary spectra of sine and cosine functions.

8. General periodic signal

Assuming $x_1(t)$ is the function of the first period of a periodic signal $x(t)$ with period of T , then:

$$\begin{aligned} x(t) &= \sum_{n=-\infty}^{\infty} x_1(t - nT) \\ &= \sum_{n=-\infty}^{\infty} x_1(t) * \delta(t - nT) = x_1(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT) = x_1(t) * \Delta_T(t) \end{aligned}$$

where

$$\Delta_T(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT)$$

The symbol “*” represents the operation of convolution, which is defined as:

$$f(t) * g(t) \equiv \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau$$

Convolution is a very important operation in signal analysis and will be introduced in more detail in following sections and chapters.

The Fourier transform of $\Delta_T(t)$ is:

$$F[\Delta_T(t)] = \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) = \omega_0 \Delta_{\omega_0}(\omega) \quad (2.55)$$

where

$$\omega_0 = \frac{2\pi}{T}$$

Therefore, according to the time domain convolution theorem (time domain convolution is equivalent to frequency domain multiplication), the Fourier transform of a general periodic signal is:

$$\begin{aligned} F(x(t)) &= F[x_1(t)] \cdot F[\Delta_T(t)] = X_1(\omega) \cdot \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) \\ &= \sum_{n=-\infty}^{\infty} [\omega_0 X_1(n\omega_0)] \cdot \delta(\omega - n\omega_0) \end{aligned} \quad (2.56)$$

$$X_n = \frac{\omega_1}{2\pi} X_0(n\omega_1) = \frac{1}{T_1} X_0(n\omega_1) \quad (2.57)$$

A comparison between Fourier transform of aperiodic signal and Fourier transform/Fourier series expansion of periodic signal is illustrated in Fig. 2.38. As we can see, the Fourier transform of an aperiodic signal is a continuous spectrum while the Fourier transform of a periodic signal is a discrete spectrum.

2.5 Properties of Fourier Transform

(1) Odd/even—real/imaginary correspondence property

If signal $x(t)$ is an even function, then imaginary part of the spectrum is zero. If the signal is an odd function, then the real part is zero.

(2) Linearity

If $x_1(t) \leftrightarrow X_1(f)$ and $x_2(t) \leftrightarrow X_2(f)$, where \leftrightarrow is a notation for Fourier transform pairs, namely $F[x_1(t)] = X_1(f)$ and $F^{-1}[X_1(f)] = x_1(t)$. c_1 and c_2 are constants, then

$$c_1 x_1(t) + c_2 x_2(t) \leftrightarrow c_1 X_1(f) + c_2 X_2(f) \quad (2.58)$$

(3) Scaling

If $x(t) \leftrightarrow X(f)$, then

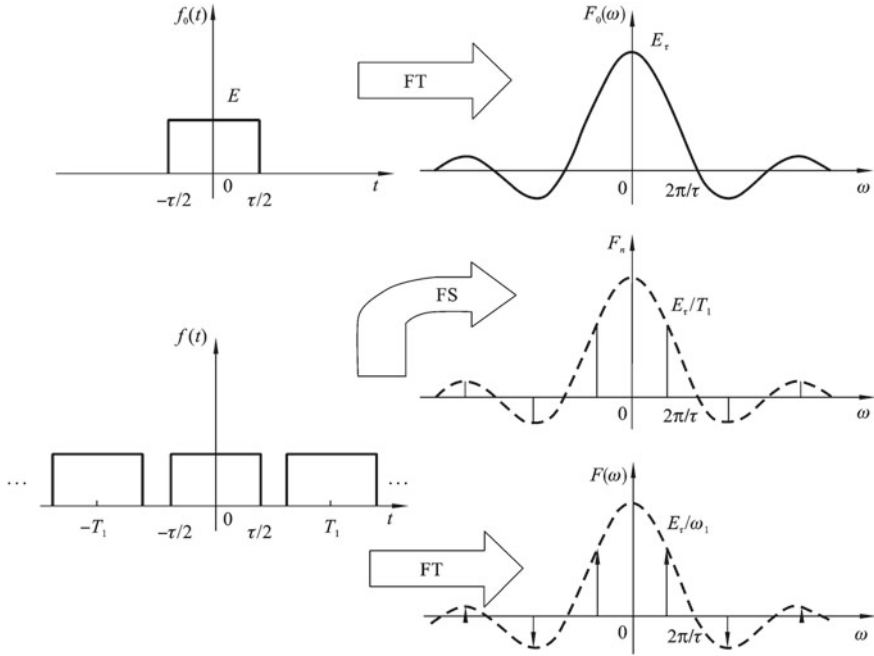


Fig. 2.38 Comparison between Fourier transform of aperiodic signal and Fourier transform/Fourier series expansion of periodic signal

$$x(kt) \leftrightarrow \frac{1}{|k|} X(f/k) \quad (2.59)$$

(4) Symmetry

If $x(t) \leftrightarrow X(f)$, then

$$X(t) \leftrightarrow x(-f) \quad (2.60)$$

(5) Time shifting

If $x(t) \leftrightarrow X(f)$, then

$$x(t \pm t_0) \leftrightarrow X(f)e^{\pm j2\pi f t_0} \quad (2.61)$$

(6) Frequency shifting

If $x(t) \leftrightarrow X(f)$, then

$$x(t)e^{\pm j2\pi f_0 t} \leftrightarrow X(f \pm f_0) \quad (2.62)$$

(7) Derivation and integration

If $x(t) \leftrightarrow X(f)$, then

$$\frac{dx(t)}{dt} \leftrightarrow j2\pi f X(f) \quad (2.63)$$

$$\int_{-\infty}^t x(t)dt \leftrightarrow \frac{1}{j2\pi f} X(f) \quad (2.64)$$

Example 2.9: Find the Fourier Transform of the Signal Shown in Fig. 2.39 The irregular pulse signal shown in Fig. 2.39 can be decomposed into the superposition of two rectangular pulse signals as shown in Fig. 2.40.

The signal in Fig. 2.39 is an even function. according to the first property, we have:

$$X_I(f) = 0$$

According to the linearity property and the Fourier transform result of single rectangular pulse in Example 2.3, we have:

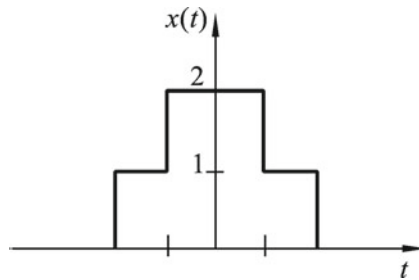


Fig. 2.39 Irregular pulse signal

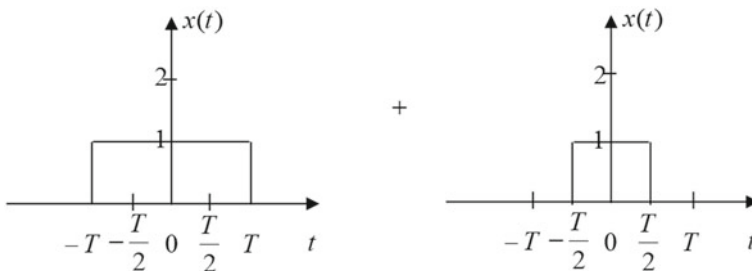


Fig. 2.40 Decomposition of the signal in Fig. 2.39

$$X_R(f) = \frac{\sin(\pi f 2T)}{\pi f} + \frac{\sin(\pi f T)}{\pi f}$$

It can be seen that the use of Fourier transform properties greatly simplifies the process of solving the problem.

2.6 Fourier Transform for Discrete-Time Signals

In engineering applications, most signals are processed by computers. Signals stored in computers are digital signals, i.e. discrete-time signals. In previous sections, we already learned Fourier series and Fourier transform for continuous time signals. In this section, we are going to learn Fourier transform for discrete-time signals. Discrete time signal is represented by a sequence $x[n]$, and the discrete-time Fourier transform (DTFT) is defined as:

$$X(f) = \sum_{n=-\infty}^{\infty} x[n] \cdot e^{-j2\pi f n}$$

Although the signal is discrete, the spectrum is continuous, it can take any frequency value ranging from negative infinity to infinity. The operation of summation goes from negative infinity to infinity, which is not applicable in computers. To make it calculable in computer, the following measures should be taken: (1) discretize the signals in both time and frequency domain; (2) limit the calculation range to a finite interval.

By discretizing the spectrum and limiting the calculation range, discrete Fourier transform (DFT) can be derived from DTFT:

$$X(k) = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi \frac{k}{N} n}$$

where k indicates the k -th frequency component. It should be noted that DTFT is different from DFT, they both take discrete time signals as input, but the first one outputs continuous frequency spectrum while the latter outputs spectrum at discrete frequencies.

There are two ways to derive the DFT expression:

- (1) Derive from the Z-transform of the discrete-time series, i.e., the discrete Fourier transform of a finite-length sequence is interpreted as its Z-transform on the unit circle;
- (2) Derive as a special case of continuous Fourier transform.

This textbook adopts the second one to show the process of derivation since it has clear physical meaning. The first method can be found in other digital signal processing books.

2.6.1 Sampling in Time and Frequency Domain

The discretization of continuous-time Fourier transform can be summarized as the steps shown in Fig. 2.41.

1. Sampling in time domain

The ideal sampling signal is a periodic impulse train $p(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_s)$. The Fourier transform of the sampling signal is

$$P(f) = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \quad (2.65)$$

For any signal $x(t)$, the signal spectrum changes before and after ideal sampling are shown in Fig. 2.42. The following conclusions can be drawn from the figure:

Conclusion 1: After the signal is sampled by periodic impulse train at the interval T_s , the Fourier transform of the sampled signal is a periodic function, which is the periodic extension of the original Fourier transform result according to the period T_s .

Conclusion 2: A discrete-time signal has periodic spectrum in frequency domain.

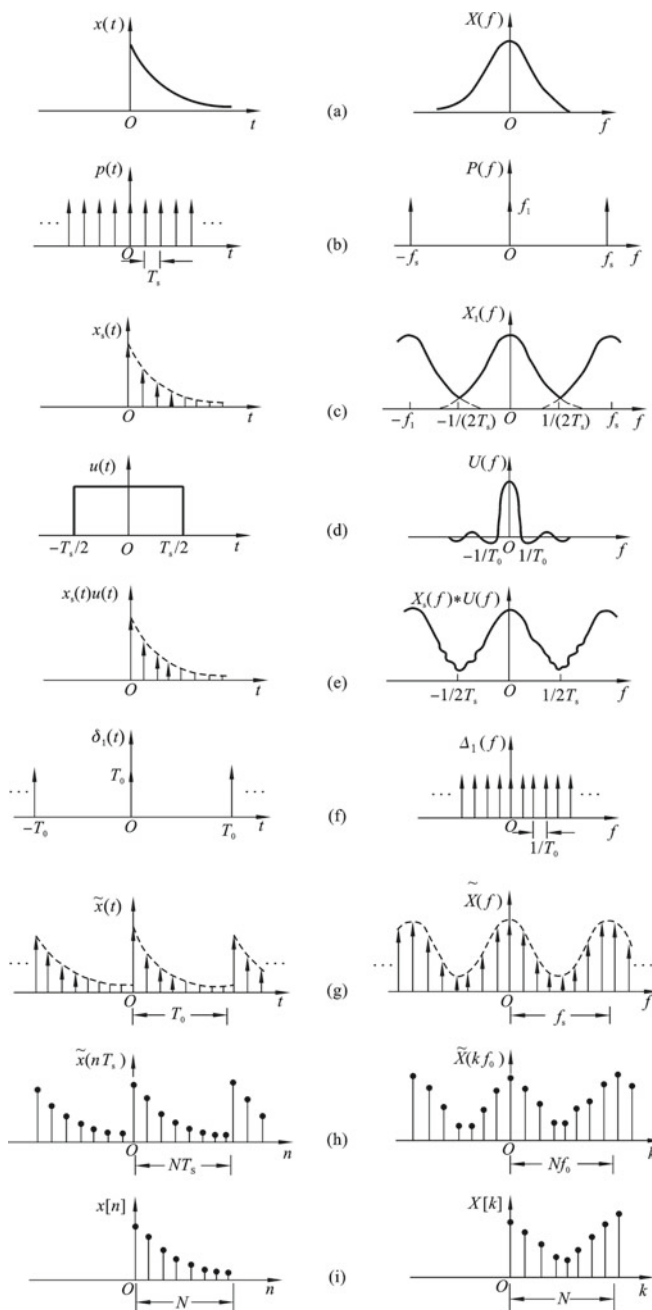
As shown in Fig. 2.42a–c, the continuous time signal $x(t)$ is sampled at the sampling interval of T_s , then the sampled signal is:

$$x_s(t) = x(t)p(t) = x(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_s) = \sum_{n=-\infty}^{\infty} x(nT_s)\delta(t - nT_s)$$

The Fourier transform of sampled signal is:

$$X_s(f) = X(f) * P(f) = f_s \sum_{n=-\infty}^{\infty} X(f - nf_s)$$

It can be seen that the frequency spectrum of the sampled signal $X_s(f)$ is a continuous periodic function, the frequency interval of the frequency spectrum is f_s , and the amplitude of the spectrum is f_s times of the spectrum $X(f)$. It can be seen from Fig. 2.42d that when the sampling interval T_s is increased, i.e. when f_s is reduced, the spectra will overlap with each other. This is the phenomenon of aliasing. To ensure that the original continuous time signal is restored without distortion from the discrete-time signal after the signal is sampled (i.e. the sampling does not cause

**Fig. 2.41** Graphical illustration of the process of discrete Fourier transform

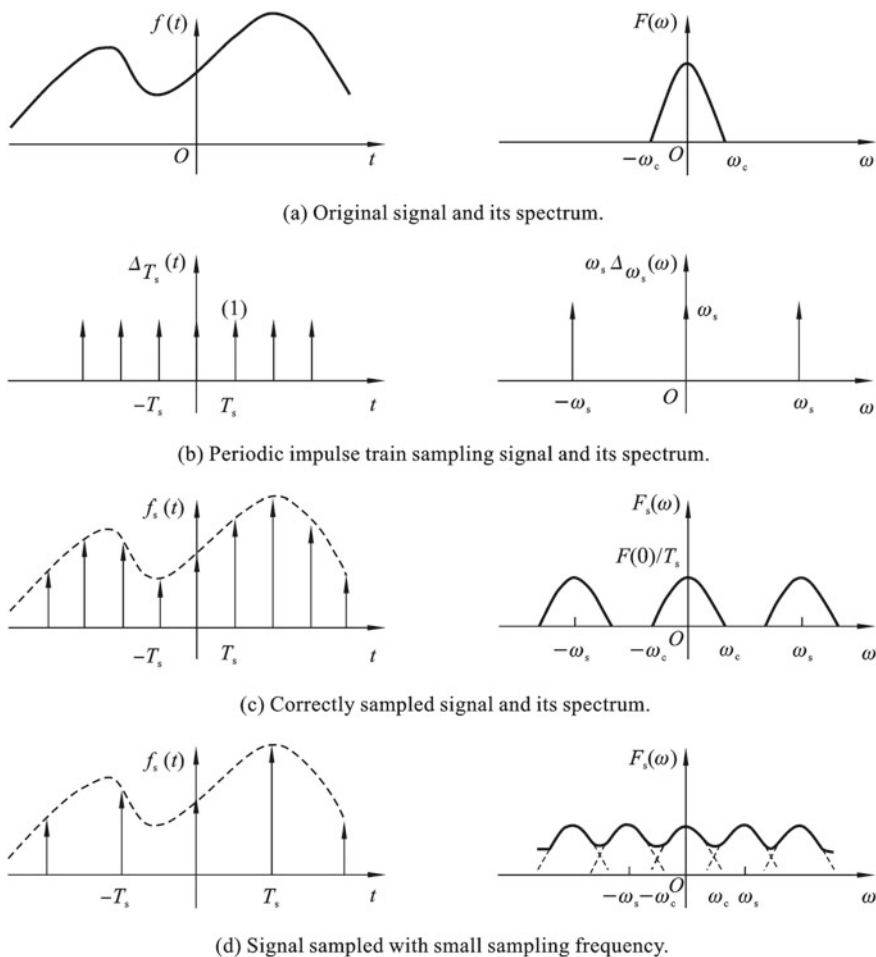


Fig. 2.42 Spectra before and after ideal sampling

any loss of information), two conditions must be met: the signal is band-limited; the sampling rate is at least twice the highest frequency of the signal.

Theoretically speaking, the method of recovering the original signal $x(t)$ from the sampled signal $x_s(t)$ is the convolution of the sampled signal and the sinc function, namely:

$$x(t) = x_s(t) * \frac{2\omega_c}{\omega_s} \text{sinc}(\omega_c t) = \frac{2\omega_c}{\omega_s} \sum_{n=-\infty}^{\infty} [x(nT_s) \text{sinc}(\omega_c(t - nT_s))] \quad (2.66)$$

The sinc function completes the interpolation operation of the discrete signal. Because of this, the sinc function is sometimes called an interpolation function. In

engineering, the method to recover the original signal from the sampled signal is to pass $x_s(t)$ through a low-pass filter with cut-off frequency of ω_c and amplification factor of T_s .

According to the sampling theorem, a continuous-time signal can be converted into a discrete-time signal, which can be further converted into a digital signal. With the development of computer technology, the theory and technology of digital signal processing become more and more important, but due to its limited data length and the process of quantization, it shows many characteristics different from analog signal processing.

2. Truncation in time domain

Use the rectangular function $u(t)$ to cut the sampled signal $x_s(t)$ so that it has only a finite number of sampling points N . Then the time function obtained after truncation is:

$$x_s(t)u(t) = \sum_{n=-\infty}^{\infty} x(nT_s)\delta(t - nT_s)u(t) = \sum_{n=0}^{N-1} x(nT_s)\delta(t - nT_s) \quad (2.67)$$

Fourier transform of sampled signal after truncation is:

$$F[x_s(t)u(t)] = X_s(f) * U(f) \quad (2.68)$$

After truncation, ripples appear in the signal spectrum as shown in Fig. 2.41e. This is caused by the truncation with rectangular window function. Because the rectangular function has a jump discontinuity, after being truncated in the time domain, ripples will appear in the frequency domain due to the Gibbs phenomenon.

After the sampled signal is truncated, it is discretized in the time domain, however it is still a continuous function in the frequency domain. In order to realize inverse transform, the frequency domain function must also be discretized.

3. Sampling in frequency domain

Let the sampling impulse train in frequency domain be $\delta_1(f)$, according to the frequency domain sampling theorem ($f_0 \leq 1/T_0$), select the sampling interval as $f_0 = 1/T_0$, where T_0 is the duration of sampled signal in time domain. And according to the symmetry property of the Fourier transform, the corresponding time domain function (Fig. 2.41f) of $\delta_1(f)$ is:

$$\delta_1(t) = T_0 \sum_{r=-\infty}^{\infty} \delta(t - rT_0)$$

The spectrum sampled by $\delta_1(f)$ is:

$$\tilde{X}(f) = X_s(f) * U(f) \cdot \delta_1(f)$$

Its inverse transform (Fig. 2.41g) is:

$$\tilde{x}(t) = x_s(t) \cdot u(t) * \delta_1(t) = T_0 \sum_{r=-\infty}^{\infty} \left[\sum_{n=0}^{N-1} x(nT_s) \delta(t - nT_s - rT_0) \right] \quad (2.69)$$

This equation shows that $\tilde{x}(t)$ is a discrete function with period T_0 , and there are N discrete points in each period. Since $\tilde{x}(t)$ is a periodic function, its Fourier transform is an equally spaced impulse train, i.e.:

$$\tilde{X}(f) = \sum_{k=-\infty}^{\infty} C_k \delta(f - kf_0) \quad k = 0, \pm 1, \pm 2, \dots$$

Its Fourier coefficients are:

$$C_k = \frac{1}{T_0} \int_{-T_s/2}^{T_0-T_s/2} \tilde{x}(t) e^{-j2\pi kt/T_0} dt \quad (2.70)$$

Substitute $\tilde{x}(t)$ inside, we get:

$$C_k = \frac{1}{T_0} \int_{-T_s/2}^{T_0-T_s/2} T_0 \sum_{r=-\infty}^{\infty} \sum_{n=0}^{N-1} x(nT_s) \delta(t - nT_s - rT_0) \cdot e^{-j2\pi kt/T_0} dt$$

The integration is carried out in a period T_0 , thus $r = 0$, and

$$\begin{aligned} C_k &= \int_{-T_s/2}^{T_0-T_s/2} \sum_{n=0}^{N-1} x(nT_s) \delta(t - nT_s) e^{-j2\pi kt/T_0} dt \\ &= \sum_{n=0}^{N-1} x(nT_s) e^{-j2\pi knT_s/T_0} \end{aligned}$$

Since $T_0 = NT_s$, we get

$$C_k = \sum_{n=0}^{N-1} x(nT_s) e^{-j2\pi kn/N} \quad (2.71)$$

and

$$\tilde{X}(f) = \sum_{k=-\infty}^{\infty} \sum_{n=0}^{N-1} x(nT_s) e^{-j2\pi kn/N} \delta(f - kf_0) \quad (2.72)$$

Re-arranging the above equations, we get:

$$\begin{cases} \tilde{x}(t) = T_0 \sum_{r=-\infty}^{\infty} \left[\sum_{n=0}^{N-1} x(nT_s) \delta(t - nT_s - rT_0) \right] \\ \tilde{X}(f) = \sum_{k=-\infty}^{\infty} \sum_{n=0}^{N-1} x(nT_s) e^{-j2\pi kn/N} \delta(f - kf_0) \end{cases} \quad (2.73)$$

This equation shows that $\tilde{x}(t)$ and $\tilde{X}(f)$ are a Fourier transform pair. It shows the time and frequency domain relationship of the signal after windowing and discretization. They are all impulse trains with N as the period, their distribution interval in time and frequency domain are both $(-\infty, \infty)$.

2.6.2 Discrete Fourier Series (DFS)

Further investigate the relationship between $\tilde{x}(t)$ and $\tilde{X}(f)$ as shown in Fig. 2.41h. The values of $\tilde{X}(f)$ is actually the Fourier coefficients C_k (Eq. 2.71) of $\tilde{x}(t)$. If we represent it by $\tilde{X}(kf_0)$, we have:

$$\tilde{X}(kf_0) = \sum_{n=0}^{N-1} x(nT_s) e^{-j2\pi kn/N} \quad k = 0, \pm 1, \pm 2, \dots \quad (2.74)$$

The sequence $\tilde{x}(t)$, as shown in Fig. 2.41g, is actually obtained by sampling the original signal $x(t)$ with N points to get $x(nT_s)$, then multiplying with a factor of T_0 . Thus it is represented as $\tilde{x}(nT_s)$. Just like $\tilde{X}(kf_0)$ is corresponding to the Fourier coefficients of $\tilde{x}(t)$, $\tilde{x}(nT_s)$ is corresponding to the Fourier coefficients of $\tilde{X}(f)$. According to the symmetry of Fourier coefficients, we have:

$$\begin{aligned} \tilde{x}(nT_s) &= \frac{1}{Nf_0} \int_0^{Nf_0} \tilde{X}(f) e^{j2\pi nT_s f} df \\ &= \frac{1}{Nf_0} \int_0^{Nf_0} \left[\sum_{k=-\infty}^{\infty} \tilde{X}(kf_0) \delta(f - kf_0) \right] e^{j2\pi nT_s f} df \\ &= \frac{1}{Nf_0} \sum_{k=0}^{N-1} \tilde{X}(kf_0) \int_0^{Nf_0} \delta(f - kf_0) e^{j2\pi nT_s f} df \\ &= \frac{1}{Nf_0} \sum_{k=0}^{N-1} \tilde{X}(kf_0) e^{j2\pi nT_s kf_0} \end{aligned}$$

$$= \frac{1}{Nf_0} \sum_{k=0}^{N-1} \tilde{X}(kf_0) e^{j2\pi nk/N} \quad n = 0, \pm 1, \pm 2, \dots \quad (2.75)$$

Thus, the following equation is obtained from Eqs. (2.74) and (2.75):

$$\begin{cases} \tilde{X}(kf_0) = \frac{1}{T_0} \sum_{n=0}^{N-1} \tilde{x}(nT_s) e^{-j2\pi nk/N} & k = 0, \pm 1, \pm 2, \dots \\ \tilde{x}(nT_s) = \frac{1}{Nf_0} \sum_{k=0}^{N-1} \tilde{X}(kf_0) e^{j2\pi nk/N} & n = 0, \pm 1, \pm 2, \dots \end{cases} \quad (2.76)$$

This equation gives the Fourier transform pair of the sampled sequence of $x(t)$. Because $\tilde{X}(kf_0)$ and $\tilde{x}(nT_s)$ are Fourier series of each other, it is usually called a discrete Fourier series (DFS) transform pair. Obviously they are sequences with N as the period, and the distribution interval in the time and frequency domain is $(-\infty, \infty)$.

2.6.3 Discrete Fourier Transform (DFT)

In Eq. (2.76), the interval of k and n ranging from 0 to $N - 1$ is defined as the “principal value interval”, and the corresponding N -point sequence in the principal value interval is defined as the “principal value sequence”. The principle value sequences are given as:

$$\begin{cases} X(kf_0) = \sum_{n=0}^{N-1} x(nT_s) e^{-j2\pi nk/N} & k = 0, 1, 2, \dots \\ x(nT_s) = \frac{1}{N} \sum_{k=0}^{N-1} X(kf_0) e^{j2\pi nk/N} & n = 0, 1, 2, \dots \end{cases}$$

Further, we use discrete sequence $x[n]$ to represent the sampled signal $x(nT_s)$, in which the n -th element in the sequence is equal to the n -th sampled signal value, $x[n] = x(nT_s)$. Similarly, we use another discrete sequence $X[k]$ to represent the sampled signal spectrum $X(kf_0)$, where $X[k]$ is the Fourier transform result at k -th frequency. Then, we obtain:

$$\begin{cases} X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} & k = 0, 1, 2, \dots \\ x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{j2\pi nk/N} & n = 0, 1, 2, \dots \end{cases} \quad (2.77)$$

As shown in Fig. 2.41i, this equation forms a discrete Fourier transform pair, which can also be expressed as:

$$\begin{cases} X[k] = \text{DFT}(x[n]) \\ x[n] = \text{IDFT}(X[k]) \end{cases} \quad (2.78)$$

Let

$$W = e^{-j2\pi/N}$$

then Eq. (2.77) becomes:

$$\begin{cases} X[k] = \sum_{n=0}^{N-1} x[n] W^{nk} & k = 0, 1, 2, \dots \\ x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] W^{-nk} & n = 0, 1, 2, \dots \end{cases} \quad (2.79)$$

The above analyses show that, through the modification of the continuous Fourier transform, the discrete time domain sequence and discrete frequency domain sequence can be connected, and the discrete Fourier transform equation is derived as shown in Eq. (2.77).

Example 2.10: Calculate Fourier Coefficients of Square Wave in Discrete Form The equation for calculating Fourier coefficients of signals in discrete form is show as follows:

$$\begin{aligned} a_n &= \frac{2}{T} \sum_{k=0}^{N-1} x(k\Delta t) \cos(2\pi n f_0 k \Delta t) \Delta t \\ b_n &= \frac{2}{T} \sum_{k=0}^{N-1} x(k\Delta t) \sin(2\pi n f_0 k \Delta t) \Delta t \end{aligned}$$

The calculation can be implemented in MATLAB with the following code, and the result is shown in Fig. 2.43.

```

Fs=5120;    N=1024; dt=1.0/Fs; T=dt*N;

t=linspace(0,T,N);

x=square(2*3.14*50*t);

subplot(2,1,1); plot(t,x,'linewidth',1);

ylim([-1.25,1.25]); grid('on');

f=linspace(0,Fs,N);

A=zeros(1,N); f0=Fs/N; cc=2*Fs/N;

for kk=1:N

    ff=(kk-1)*f0;  am=0; bm=0 ;

    for k=1:N

        am=am+x(k)*cos(2*pi*ff*k*dt)*dt;

        bm=bm+x(k)*sin(2*pi*ff*k*dt)*dt ;

    end

    re=cc*am; im=cc*bm;

    A(kk)=sqrt(re*re+im*im);

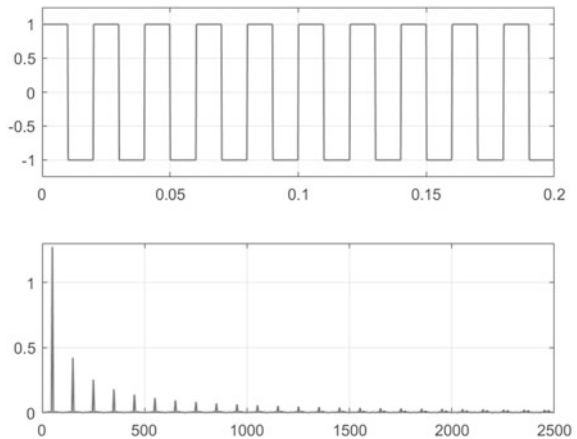
end

subplot(2,1,2); plot(f,A,'linewidth',1);

ylim([0,1.3]); xlim([0,2500]); grid('on');

```

Fig. 2.43 Fourier coefficients of square wave



2.6.4 Fast Fourier Transform (FFT)

Fast Fourier transform is an algorithm to reduce the calculation time of DFT. Although DFT is a feasible transformation tool for the analysis of discrete signals, it is difficult to implement because of the long calculation time. For example, for the sampling point $N = 1000$, the DFT algorithm requires about 2 million operations, while the FFT only requires about 15 thousand times. It can be seen that the FFT algorithm greatly improves the operation efficiency. When the FFT algorithm was first proposed by James W. Cooley and John W. Tukey in 1965, it was considered an epoch-making advancement in signal analysis technology.

There are many variants of the FFT algorithm, and the establishment of each variant mainly takes into account the characteristics of the analyzed data and the computer characteristics. The typical form of FFT is the Cooley-Tukey algorithm. Because it decomposes the time series $x[n]$, it is often referred as a method of time decimation. The other algorithm is Sande-Tukey algorithm, which decomposes $X[k]$ in frequency domain, so it is often referred as a method of frequency decimation.

The FFT algorithm is still under development, new FFT algorithms continue to appear. For example, the NFFT algorithm proposed by Rader; the WFTA algorithm proposed by Winograd; the PFTA algorithm proposed by Nussbaumer. All those algorithms improve the calculation speed in some extent. When $N = 1000$, the multiplication operations in each algorithm are: DFT—2 million; FFT—15 thousand; NFFT—8 thousand; WFTA—3.5 thousand; PETA—3 thousand.

Example 2.11: Demonstration of Operations in DFT Take the coefficient a_i as an example:

$$\begin{aligned}
 a_1 &= x(0) \cos(2\pi f_0 0 \Delta t) + x(1) \cos(2\pi f_0 1 \Delta t) + x(2) \cos(2\pi f_0 2 \Delta t) + x(3) \cos(2\pi f_0 3 \Delta t) + \dots \\
 a_2 &= x(0) \cos(2\pi 2 f_0 0 \Delta t) + x(1) \cos(2\pi 2 f_0 1 \Delta t) + x(2) \cos(2\pi 2 f_0 2 \Delta t) + x(3) \cos(2\pi 2 f_0 3 \Delta t) + \dots \\
 a_3 &= x(0) \cos(2\pi 3 f_0 0 \Delta t) + x(1) \cos(2\pi 3 f_0 1 \Delta t) + x(2) \cos(2\pi 3 f_0 2 \Delta t) + x(3) \cos(2\pi 3 f_0 3 \Delta t) + \dots \\
 &\dots \\
 a_n &= x(0) \cos(2\pi n f_0 0 \Delta t) + \dots
 \end{aligned}$$

We can see that there are a lot of operations of sine/cosine. FFT is an efficient way of calculating DFT, it reduces the amount of calculation by selecting and arranging intermediate results. DFT algorithm responds 100 times within 9 s, while FFT algorithm responds 50,000 times within 1 s.

Example 2.12: Calculating Magnitude and Phase Spectra with FFT The FFT algorithm can be implemented in MATLAB easily. The following code is an example of spectrum analysis of a signal with two frequencies (Fig. 2.44).

```

Fs=5120;    N=1024;

dt=1.0/Fs; T=dt*N;

t=linspace(0,T,N);

x=10*sin(2*3.14*100*t)+3*sin(3*2*3.14*100*t);

subplot(3,1,1); plot(t,x);

y=fft(x,N);

% N/2, Amplitude be corrected

A1=abs(y)/(N/2);

Q1=angle(y)*180/pi;

%f, x axis

f=linspace(0,Fs/2,N/2);

subplot(3,1,2);

%0-N/2, only positive frequency

plot(f,A1(1:N/2));

subplot(3,1,3);

plot(f,Q1(1:N/2));

```

Example 2.13: Calculating Power Spectrum with FFT The power spectrum can be similarly calculated in MATLAB. We only need one more line to calculate the power from Fourier transform result, as shown in Fig. 2.45.

Fig. 2.44 Magnitude and phase spectra calculated using FFT

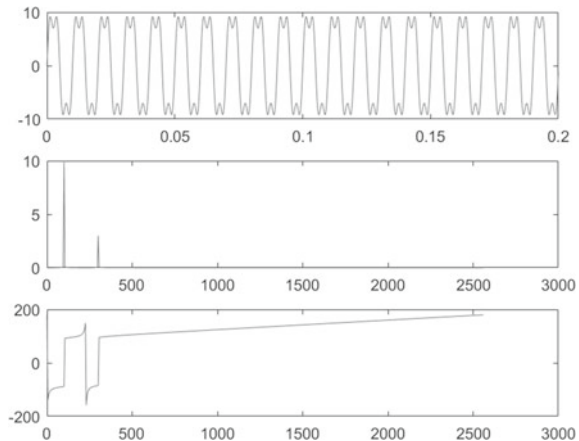
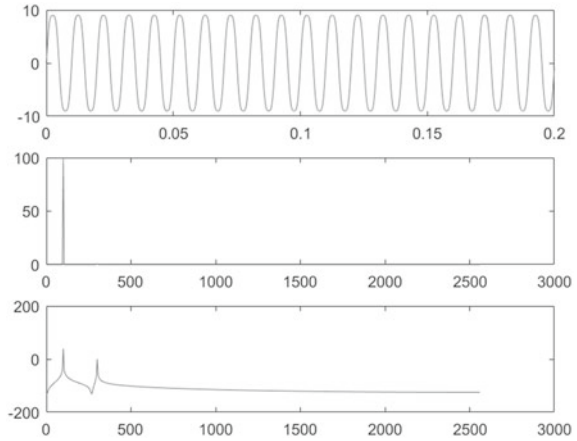


Fig. 2.45 Power spectrum calculated by FFT



```

Fs=5120; N=1024;

dt=1.0/5120.0; T=dt*N;

t=linspace(0,T,N);

x=10*sin(2*3.14*100*t)+sin(3*2*3.14*100*t);

subplot(3,1,1); plot(t,x);

y=fft(x,N);

f=linspace(0,Fs/2,N/2);

A1=abs(y)/(N/2);

A2=A1.^2;

P2 =20*log10(A2);

subplot(3,1,2); plot(f,A2(1:N/2));

subplot(3,1,3); plot(f,P2(1:N/2));

```

2.6.5 Applications of FFT Algorithm

From the above analysis, we can see that FFT is a fast algorithm for implementing DFT. DFT can be considered as a combination of a periodic discrete time sequence and a periodic discrete frequency domain sequence. They are related through a process of finite summation. This summation process can be completed by using

a computer with the help of an FFT algorithm, which provides a fast spectrum analysis method. The FFT algorithm can be used directly to process discrete signals, and can also be used to approximately process continuous-time signal. The specific applications of FFT algorithm mainly include approximating the result of Fourier transform, harmonic analysis, fast convolution operation, fast correlation calculation and power spectrum estimation.

1. Approximate the result of Fourier transform

A comparison of Fourier transform and discrete Fourier transform for a one-sided exponential function $x(t) = e^{-1}u(t)$ is shown in Fig. 2.46. The solid line shows the continuous Fourier transform result, and the dotted line shows the DFT result. It can be seen that the DFT is an approximation of FT. Its real part and imaginary parts are respectively even and odd functions that are symmetrical about the point $k = N/2$. When $k > N/2$, the spectra show the result of negative frequency. As we can see, there is a good approximation at the low frequency part, while having large error at the high frequency part. This is caused by the spectral leakage during frequency aliasing and truncation.

The DFT result is obtained via the following process:

- (1) Sampling in time domain with $N = 32$, $T = 0.25$ s to get the sequence $x[n]$;
- (2) Apply the FFT algorithm:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi nk}$$

$X[k]$ is a complex number with real part $X_R(k)$ and imaginary part $X_I(k)$:

$$X[k] = X_R[k] + jX_I[k]$$

The magnitude and phase spectra can be derived as:

$$|X[k]| = \sqrt{X_R^2[k] + X_I^2[k]}$$

$$\varphi[k] = \arctan \frac{X_I[k]}{X_R[k]}$$

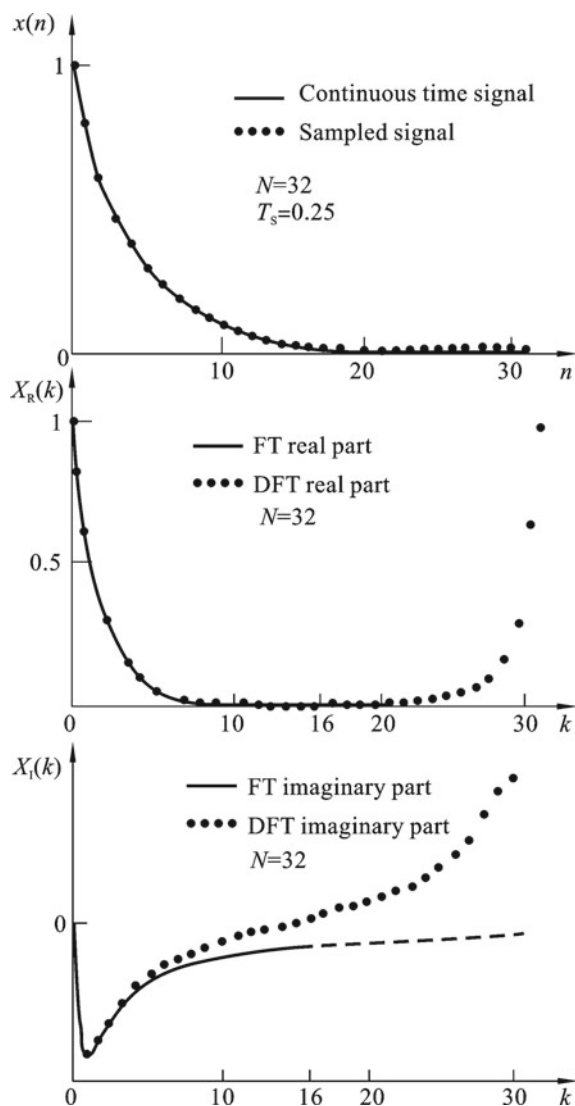
2. Harmonic analysis

To use DFT for harmonic analysis of periodic square wave, we need to calculate:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi nk/N}$$

to get the coefficient of each harmonics. As shown in Fig. 2.47, the square wave is sampled with sampling points of $N = 32$ and the spectrum is calculated by FFT

Fig. 2.46 Comparison of DFT and FT for one-sided exponential function

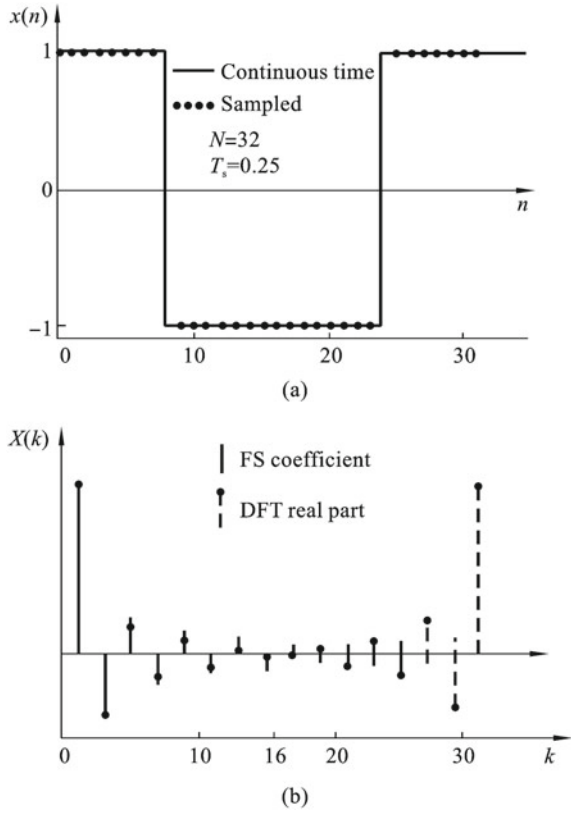


algorithm. The resulting spectrum is symmetrical about $k = N/2$. The low-order harmonics are relatively well fitted, while the high-order harmonics have errors. This is due to the frequency aliasing effect, and the errors can be reduced by increasing the sampling rate.

3. Fast convolution operation

The FFT algorithm can realize the calculation of discrete convolution. Discrete convolution has the same meaning as the convolution of a continuous time system.

Fig. 2.47 Harmonic analysis of periodic square wave



When describing the output and input relationship of a discrete time system, the output $y[n]$ is the convolution of the input $x[n]$ and the system unit impulse response $h[n]$:

$$y[n] = \sum_{m=-\infty}^{\infty} x[m]h[n-m] = x[n] * h[n]$$

or

$$y[n] = \sum_{m=-\infty}^{\infty} h[m]x[n-m] = h[n] * x[n] \quad (2.80)$$

The operation process includes flipping, translation, multiplication and summation. As shown in Fig. 2.48, $y[n]$ is the convolution of two sequences $x[n]$ and $h[n]$. The sequence $y[n]$ is calculated as:

$$y[0] = 1 \times 4 = 4$$

$$\begin{aligned}
 y[1] &= 1 \times 3 + 2 \times 4 = 11 \\
 y[2] &= 1 \times 2 + 2 \times 3 + 3 \times 4 = 20 \\
 &\vdots
 \end{aligned}$$

It can be seen that the convolution operation process is complicate. When the number of sampling points of the sequence $x[n]$ and $h[n]$ are N_1 and N_2 respectively, the number of multiplication operations is $N_1 \times N_2$. Obviously, when the number of

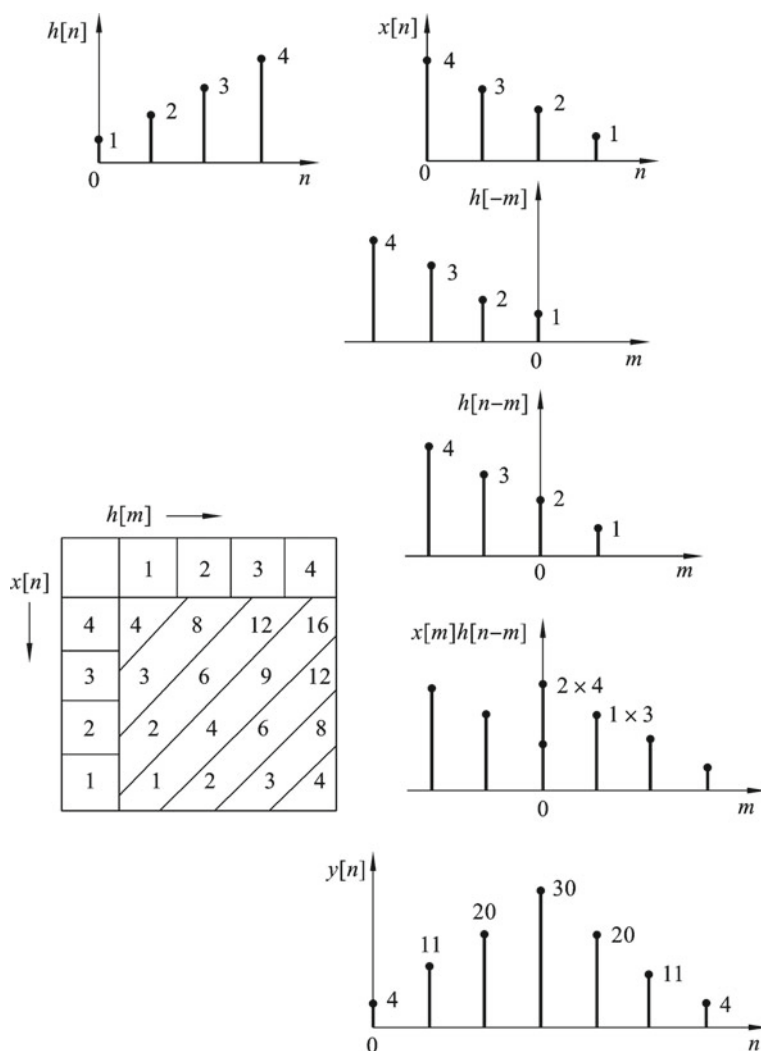


Fig. 2.48 Process of calculating convolution of two sequences

sampling points is large, even if it is calculated on a computer, there will be difficulties in either the calculation time or the amount of storage required for the calculation.

According to the time domain convolution theorem of the Fourier transform, there will be two functions with a period of N satisfying:

$$\sum_{m=0}^{N-1} x[m]h[n-m] \Leftrightarrow X[k]H[k]$$

or

$$x[n] * h[n] \Leftrightarrow X[k]H[k] \quad (2.81)$$

Based on this theorem, to calculate the convolution of two time domain sequences $x[n]$ and $h[n]$, we can calculate their DFTs $X[k]$ and $H[k]$ separately, multiply them and apply the inverse DFT to get the convolution result. This theorem provides the basis for calculating time domain convolution with fast Fourier transform.

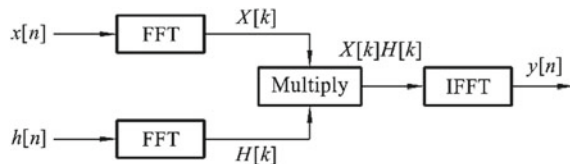
The process of implementing fast convolution is shown in Fig. 2.49, which includes three steps: (1) use FFT algorithm to calculate DFT of two signals; (2) multiply the transformed values of the two signals at each frequency point; (3) use the IFFT algorithm to calculate the inverse transform of the multiplied result. A total of two FFT and one IFFT operations are required to implement this process, which is equivalent to three FFT operations. In the design of finite impulse response (FIR) digital filters, the step of obtaining $H[k]$ from $h[n]$ is pre-implemented, and the $H[k]$ result has been stored in the memory, so only two FFT operations are actually required. Let $N_1 = N_2 = N$, then $2(\frac{N}{2} \log_2 N)$ times of multiplication operations are required. Besides, there are N times of multiplication operations for the multiplication of $X[k]$ and $H[k]$. Thus, the total number of operations is:

$$2\left(\frac{N}{2} \log_2 N\right) + N = N(1 + \log_2 N)$$

Obviously, as N increases, this number is significantly less than N^2 (the operations required for direct calculation of convolution).

It should be noted that in the implementation of the fast convolution algorithm, due to the use of DFT analysis, both time domain and frequency domain are periodic discrete sequences, when they are called in the periodic convolution, there will be an iterative summation between periodic data, which brings a so-called wraparound

Fig. 2.49 Algorithm of fast convolution



error to the calculation result. The way to avoid the wraparound error is to fill in zeros at the end of $x[n]$ and $h[n]$. Usually, we double the period. If $x[n]$ and $h[n]$ are equal in length, both are lengthened by N points.

The fast convolution process can be carried out as follows:

- (1) Use zero padding to modify $x[n]$ and $h[n]$ to avoid wraparound errors;
- (2) Use the FFT algorithm to calculate the DFT of the two modified functions, get $X[k]$ and $H[k]$;
- (3) Multiply $X[k]$ and $H[k]$ to get:

$$Y[k] = X[k] \cdot H[k]$$

- (4) Use FFT algorithm to calculate the IDFT, namely

$$y[n] = \text{IDFT}(Y[k])$$

The $y[n]$ is the computed convolution of $x[n]$ and $h[n]$.

4. Calculation of correlation

There are two calculation methods for correlation: one is direct calculation in time domain and the other is the FFT algorithm. The cross-correlation of two signals x and y is defined as:

$$r_{xy}[m] = \frac{1}{N} \sum_{n=0}^{N-1} x[n]y[n-m] \quad (2.82)$$

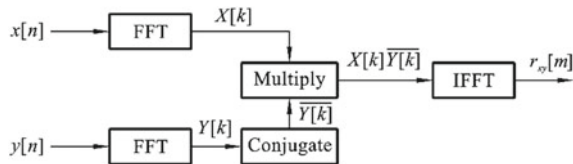
And the auto-correlation of a signal x is defined as:

$$r_x[m] = \frac{1}{N} \sum_{n=0}^{N-1} x[n]x[n-m] \quad (2.83)$$

This operation is similar to the convolution, and it also includes multiplications and summations of sequences, which requires a lot of calculation time.

The FFT algorithm of correlation is based on the Wiener–Khinchin relationship, that is, the autocorrelation function or the cross-correlation function can be obtained from the power spectral density or the cross-spectral density function. The steps of calculation are shown in Fig. 2.50.

Fig. 2.50 Steps of calculating correlation by FFT algorithm



The steps are:

- (1) Apply the FFT algorithm to $x[n]$ and $y[n]$ to get their spectra $X[k]$ and $Y[k]$;
- (2) Take the conjugation of one spectrum and multiply them:

$$S_{xy}[k] = \frac{X[k]\overline{Y[k]}}{N}$$

Or, when $y[n]$ is the same as $x[n]$, take the auto power spectral density:

$$S_x[k] = \frac{1}{N}|X[k]|^2$$

- (3) Apply IFFT and get correlation from the power spectral density:

$$\begin{aligned} r_{xy}[m] &= \text{IFFT}(S_{xy}[k]) \\ r_x[m] &= \text{IFFT}(S_x[k]) \end{aligned}$$

Although this is an indirect method, it is 5–100 times faster than the direct time domain calculation method.

When using the FFT method to calculate the correlation function, we must also pay attention to the influence of the wraparound error, which is similar to the error in the fast convolution. The solution is also to extend the time sequences $x[n]$ and $y[n]$ with zero padding.

5. Power spectral density analysis

Traditional spectral analysis methods are based on Fourier transform analysis methods, including correlation function method and periodogram method. The correlation function method is a practical algorithm proposed by Blackman and Tukey in 1958, which is also known as the BT method. It uses statistical analysis to obtain the autocorrelation function of the signal in time domain, and then perform the Fourier transform to obtain the power spectrum; the periodogram method directly processes the data with Fourier transform, and then takes the square of the amplitude to get the power spectral density of the signal. The periodogram method was used in spectrum estimation after the FFT method proposed by Cooley-Tukey in 1965.

Usually, the value of a random signal at any point in time axis is not a priori, and the samples are always different, so it cannot be accurately represented by mathematical formulas or graphs, but can only be represented by various statistical parameters. Among them, the autocorrelation is the parameter that can comprehensively characterize the statistical averages of a random signal. And, the power spectral density of a random signal is the Fourier transform of the autocorrelation function. For random signals, its spectrum does not exist, and only the power spectral density can be used to characterize its statistical spectral characteristics. Therefore, power spectral density is one of the most important characterization forms of random signals. To understand a random signal in a statistical sense, it is necessary to know its power spectral density.

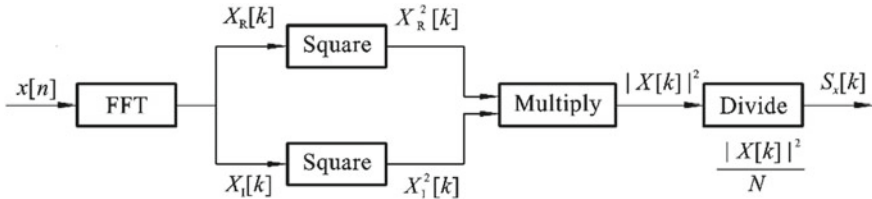


Fig. 2.51 Steps of calculating power spectral density using periodogram

According to the theory of autocorrelation and Wiener–Khinchin relationship, it is easy to prove that the power spectral density of a random sequence $x[n]$ is:

$$S_x[k] = \lim_{N \rightarrow \infty} \frac{1}{N} |X[k]|^2 \quad (2.84)$$

Its estimation is:

$$\hat{S}_x[k] = \frac{1}{N} |X[k]|^2 \quad (2.85)$$

where

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N} \quad k = 0, 1, \dots, N-1$$

If we can observe N values of $x[n]$, then we can use FFT to calculate $X[k]$ and use Eq. (2.85) to get the estimation of $S_x[k]$. The steps of calculating power spectral density is shown in Fig. 2.51.

Since the discrete Fourier transform $X[k]$ of the sequence $x[n]$ is periodic, it is called the periodogram of a real stationary random signal sequence $x[n]$. Obviously, to express the power spectrum of a random signal with a periodogram, it is only necessary to perform a DFT operation on the signal sequence itself, then take the square of its absolute value, and divide the length of the sequence. Therefore, using the FFT algorithm, it is easy to directly estimate the approximate value of the power spectral density of a real random sequence.

The analysis shows that when using the periodogram method for spectrum estimation, there are two main problems, namely the statistical variability of the power spectral density and the problem of spectral leakage. The former is a statistical error due to the uncertainty caused by limited amount of data collected in the power spectrum measurement; while the latter is inherent in the spectrum analysis, it will cause the skewness error of the estimation. Therefore, in practice, measures are often taken to modify the periodogram method to minimize the estimation error. The measures are: (1) adopting averaging to reduce statistical variability; (2) adopting windowing processing to reduce leakage.

The above two traditional methods are essentially the same, and both regard a finite-length data segment as the result of windowing and truncation of an infinitely long sampling sequence. Whether it is windowing of data or windowing of auto-correlation function, the phenomenon of “leakage” occurs in the frequency domain, i.e. the energy in the main lobe of the power spectrum leaks into the side lobes, so that the main lobe of a weak signal is easily affected by the side lobes of a strong signal. It causes blur and distortion of the frequency spectrum. In order to reduce the side lobes, many researchers try to modify the windowing function, but all side lobe suppression techniques are at the expense of spectral resolution. However, in spectrum analysis applications, frequency resolution is as important as low side lobes, and sometimes even more important. Thus, solving the contradiction between high frequency resolution and low side lobes becomes a major issue in spectrum analysis. In addition, with the traditional spectrum estimation method, only when the number of sampling points is large, can the spectrum estimation accuracy be higher. This will not only increase the work of data processing, but also cannot do anything about the short data or transient signals in engineering application and research. Under this background, modern spectrum analysis methods, maximum entropy spectrum analysis has been proposed.

2.7 Error Analysis in FFT

2.7.1 Signal Truncation

Fourier transform is the main mathematical tool for studying relationship between time domain and frequency domain in digital signal processing. However, when using a computer to process measurement signals, it is impossible to perform calculations on infinitely long signals. Instead, it takes a limited time interval (time window) for analysis, which requires truncation. The truncation method is to multiply the infinite signal by the window function. The term “window” here means that a part of the signal can be observed through a window, and the rest is shielded (treated as zero). As shown in Fig. 2.52, cosine signal is distributed in $(-\infty, \infty)$. When it is multiplied with the window function $w(t)$, the truncated signal $x_T(t) = x(t)w(t)$ with finite length is obtained. According to the Fourier transform relationships, the spectrum of the cosine signal $X(f)$ is a δ function at f_0 , and the spectrum of the rectangular window function $w(t)$ is the $\text{sinc}(f)$ function. According to the frequency domain convolution theorem, the spectrum of truncated signal $x_T(t)$ is:

$$X_T(f) = \frac{1}{2\pi} X(f) * W(f)$$

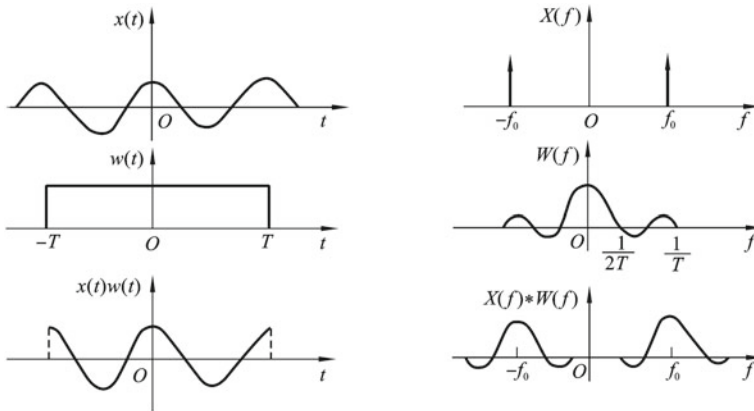


Fig. 2.52 Truncation and spectral leakage of cosine signal

2.7.2 Effect of Spectral Leakage

Comparing the spectrum of the truncated signal $X_T(f)$ with the spectrum of the original signal $X(f)$, it can be seen that it is no longer two spectral lines, but the continuous spectrum of two oscillations. This shows that after the original signal is truncated, its frequency spectrum has been distorted, and the energy originally concentrated at f_0 is dispersed into nearby frequencies. This phenomenon is called spectral leakage.

The spectral leakage phenomenon caused by signal truncation is inevitable, because the window function $w(t)$ is a function with infinite frequency band, so even if the original signal $x(t)$ is a band-limited signal, it will inevitably become infinite after truncation, i.e. the distribution of energy in the frequency domain is expanded. It is also known from the sampling theorem that no matter how high the sampling frequency is, as long as the signal is truncated, it will inevitably cause aliasing. Therefore, signal truncation will inevitably lead to some errors, which is a problem that cannot be ignored in signal analysis.

If the truncation length T is increased, i.e. the rectangular window is widened, its spectrum $W(f)$ will be compressed and narrowed since $1/(2T)$ decreases. Although theoretically its spectral range is still infinitely wide, in fact the frequency components outside the center frequency attenuate fast, so the leakage error will be reduced. When the window width T tends to infinity, the window spectrum $W(f)$ will become a $\delta(f)$ function, and the convolution of $\delta(f)$ and $X(f)$ will still be $X(f)$. This infers that if the window is infinitely wide, i.e. it is not truncated, there is no leakage error.

Leakage is related to the side lobes on both sides of the window function spectrum. If the height of the side lobe tends to zero, and the energy is relatively concentrated in the main lobe, it can be closer to the true frequency spectrum. For this reason, different window functions can be used to cut off the signal in the time domain.

2.7.3 Fence Effect

As we have introduced in previous sections, in order to calculate and display spectrum in a computer, we have to discretize the spectrum. If the peak of spectrum does not coincide with the sampled frequency, we can only take the spectrum line value of the adjacent frequency instead as shown in Fig. 2.53. This phenomenon is called fence effect.

Example 2.14: Demonstration of Fence Effect in MATLAB The fence effect can be demonstrated with the following MATLAB code. The total duration of signal is 0.2 S, which makes the sampling period in frequency to be $\Delta f = 5$ Hz. From the spectrum in Fig. 2.54, we can observe that the magnitude of the 500 Hz signal matches the signal amplitude. Whereas for the signal of 102 Hz, there is error in the magnitude.

```

Fs=5120; N=1024;

dt=1.0/5120.0; T=dt*N; %df=5Hz

t=linspace(0,T,N);

x1=sin(2*3.14*102*t);

x2=sin(2*3.14*500*t);

x=x1+x2;

subplot(2,1,1); plot(t,x,'linewidth',1);

xlim([0,0.1]); ylim([-2.5,2.5]); grid('on');

f=linspace(0,Fs,N);

y=fft(x,N);

A1=abs(y)/(N/2);

subplot(2,1,2); plot(f,A1,'linewidth',1);

ylim([0,1.25]); xlim([0,2500]); grid('on');
```

Fig. 2.53 Fence effect

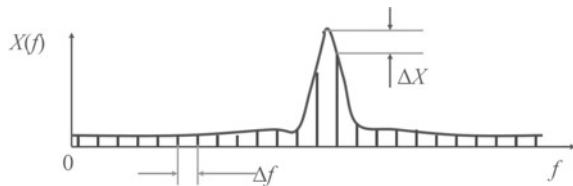
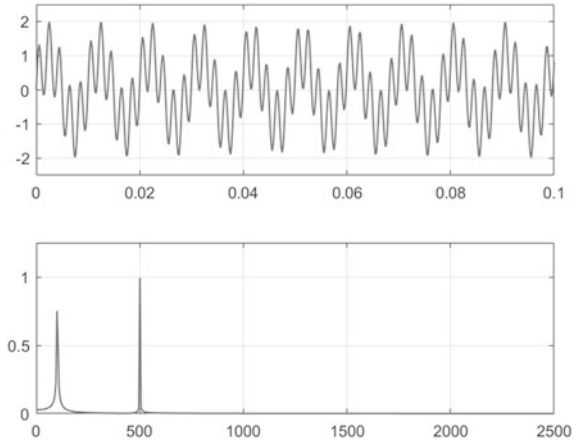


Fig. 2.54 Demonstration of fence effect



2.7.4 Window Functions

The commonly used window functions in signal processing are: rectangular window, triangular window, trapezoid window, sine window, Hann window, Hamming window, Gaussian window, etc. The properties of these windows will be introduced.

(1) Rectangular window. It is defined as:

$$w(t) = \begin{cases} \frac{1}{T}, & |t| \leq T \\ 0, & |t| > T \end{cases} \quad (2.86)$$

Its spectrum is:

$$W(f) = \frac{\sin 2\pi f T}{\pi f T} \quad (2.87)$$

Rectangular window is the most widely used window. Customarily, if we say add window, the window is default to be a rectangular window. The advantage of this kind of window is that the main lobe is relatively concentrated, but the disadvantage is that the side lobes are high and there are negative side lobes (Fig. 2.55), which leads to high-frequency interference and leakage in the transformation.

(2) Triangular window. It is also called Fejér window, and is defined as:

$$\omega(t) = \begin{cases} \frac{1}{T} \left(1 - \frac{|t|}{T}\right), & |t| \leq T \\ 0, & |t| > T \end{cases} \quad (2.88)$$

Its spectrum is:

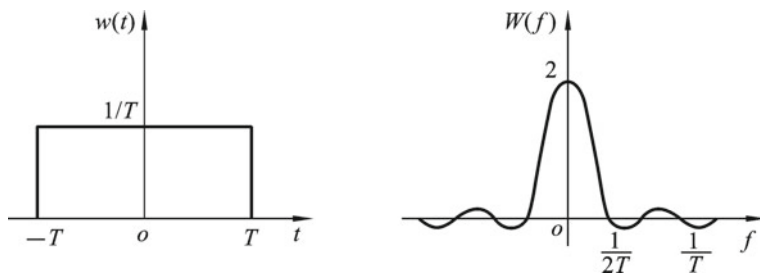


Fig. 2.55 Rectangular window and its spectrum

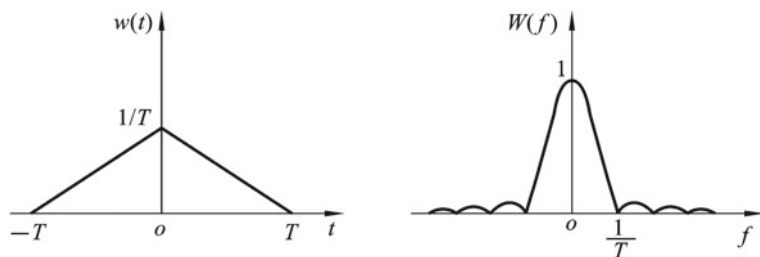


Fig. 2.56 Triangular window and its spectrum

$$W(f) = \left(\frac{2 \sin \pi f T}{\pi f T} \right)^2 \quad (2.89)$$

Comparing with the rectangular window, the width of main lobe of triangular window is approximately twice that of the rectangular window, but the side lobes are small and there are no negative side lobes, as shown in Fig. 2.56.

- (3) Hann window. It is named after Julius von Hann and also called raised cosine window. It is defined as:

$$W(t) = \begin{cases} \frac{1}{T} \left(\frac{1}{2} + \frac{1}{2} \cos \frac{\pi t}{T} \right), & |t| \leq T \\ 0, & |t| > T \end{cases} \quad (2.90)$$

Its spectrum is:

$$W(f) = \frac{\sin 2\pi f T}{2\pi f T} + \frac{1}{2} \left[\frac{\sin(2\pi f T + \pi)}{2\pi f T + \pi} + \frac{\sin(2\pi f T - \pi)}{2\pi f T - \pi} \right] \quad (2.91)$$

It can be seen that the spectrum of Hann window can be regarded as the sum of the spectra of three rectangular windows, or the sum of three sinc functions. The two terms in the brackets can be regarded as the first term moved towards left and right by $1/2T$. Therefore, the side lobes cancel each other and the high-frequency

interference and leakage are reduced. The frequency spectrum of the Hann window is shown in Fig. 2.57.

A comparison between Hann window and rectangular window is shown in Fig. 2.58. In which, Fig. 2.58a shows the relationship between $W(f)$ and f , Fig. 2.58b shows the relationship between $\log f$ and the decay of amplitude with respect to the main lobe. It can be seen that the main lobe of the Hann window is widened (the first zero-crossing point is at $1/T$) and the side lobe is significantly reduced. The first side lobe is attenuated by -32 dB for the Hann window, while it is attenuated by -13 dB for the rectangular window. In addition, the side lobe attenuation speed of the Hann window is also faster, which is about 60 dB/(10 oct), while that for rectangular window is 20 dB/(10 oct). From the above comparison, it can be inferred that the Hann window is better than the rectangular window from the perspective of reducing leakage. But the main lobe of the Hann window is widened, which is equivalent to widening the analysis bandwidth and decreasing the frequency resolution.

- (4) Hamming window. It is named after Richard W. Hamming, and is also a type of cosine window. It is defined as:

$$\omega(t) = \begin{cases} \frac{1}{T}(0.54 + 0.46 \cos \frac{\pi t}{T}), & |t| \leq T \\ 0, & |t| > T \end{cases} \quad (2.92)$$

Its spectrum is:

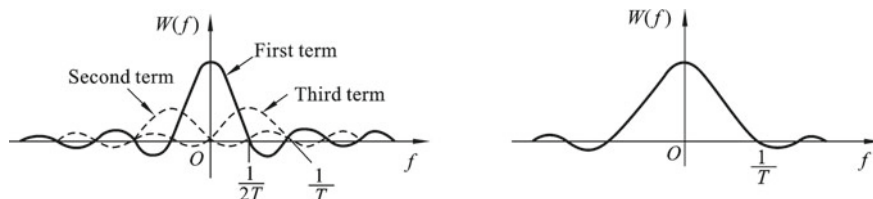


Fig. 2.57 Hann window and its spectrum

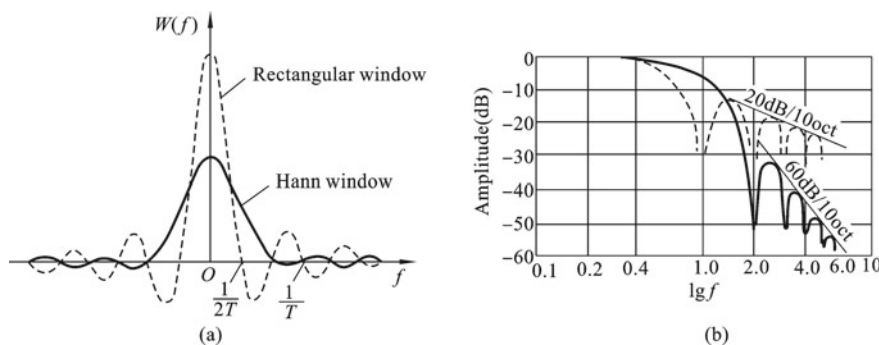


Fig. 2.58 Comparison between Hann window and rectangular window

$$W(f) = 1.08 \frac{\sin 2\pi f T}{2\pi f T} + 0.46 \left[\frac{\sin(2\pi f T + \pi)}{2\pi f T + \pi} + \frac{\sin(2\pi f T - \pi)}{2\pi f T - \pi} \right] \quad (2.93)$$

Hamming window and the Hann window are cosine windows with different weighting coefficients. The weighted coefficient of the Hamming window makes the side lobes smaller. Analysis shows that the first side lobe attenuation of the Hamming window is -42 dB. The frequency spectrum of Hamming window is also the sum of three rectangular window spectra, but the attenuation speed is 20 dB/(10 oct), which is slower than the Hann window. Both Hamming window and Hann window are very widely used.

(5) Gaussian window. It is a type of exponential window, and is defined as:

$$\omega(t) = \begin{cases} \frac{1}{T} e^{-\alpha t^2}, & |t| \leq T \\ 0, & |t| > T \end{cases} \quad (2.94)$$

where α is a constant which decides how fast the function curve decays. If α is selected properly, the function value at the truncation point is relatively small, and the truncation will cause little effect to the signal. The spectrum of Gaussian window does not have negative side lobes, and the first side lobe attenuation is -55 dB. The main lobe is relatively wide, so the frequency resolution is low. The Gaussian window function is often used to cut off some aperiodic signals, such as exponentially decaying signals.

In addition to the above window functions, there are also some commonly used window functions, such as Parzen window, Blackman window, Kaiser window and so on. Table 2.1 lists the characteristics of five typical window functions.

Regarding the selection of the window function, the property of the analyzed signal and the requirements of process should be considered. If only the frequency resolution is concerned, and the magnitude of each frequency of not very important, a rectangular window with a narrower main lobe width can be selected. For example, to process the signal of vibration and to find its natural frequency, the rectangular window should be selected. If the signal to be analyzed has narrow bandwidth and large noise, a window function with a small side lobe amplitude should be selected,

Table 2.1 Characteristics of typical window functions

Window function	-3 dB bandwidth	Equivalent noise bandwidth	Side lobe amplitude (dB)	Side lobe attenuation speed (dB/(10 oct))
Rectangular	0.89 B	B	-13	-20
Triangular	1.28 B	1.33 B	-27	-60
Hann	1.20 B	1.23 B	-32	-60
Hamming	1.30 B	1.36 B	-42	-20
Gaussian	1.55 B	1.64 B	-55	-20

such as Hann window and triangular window. For functions that decay exponentially with time, an exponential window can be used to improve the signal-to-noise ratio.

Example 2.15: Analyze Window Functions with MATLAB The following MATLAB code shows the process of windowing. It calculates the time domain waveform and frequency spectrum of the windowed signal. The result is shown in Fig. 2.59.

```

Fs=5120; N=1024; dt=1.0/5120.0;

T=dt*N; t=linspace(0,T,N);

x=10*sin(2*3.14*102*t);

subplot(5,1,1);

plot(t,x,'linewidth',1);

f=linspace(0,Fs/2,N/2);

y=fft(x,N); A1=abs(y)/(N/2);

subplot(5,1,2);

plot(f,A1(1:N/2),'linewidth',1);

w=flattopwin(N); w1=w';

subplot(5,1,3);

plot(t,w1,'linewidth',1);

z=w1.*x;

subplot(5,1,4);

plot(t,z,'linewidth',1);

y1=fft(z,N); A2=4.545*abs(y1)/(N/2);

subplot(5,1,5);

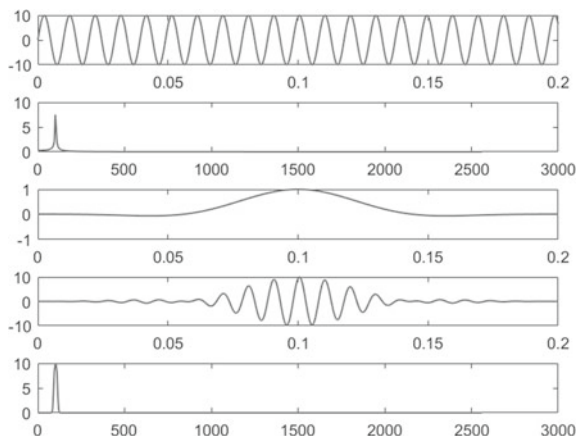
plot(f,A2(1:N/2),'linewidth',1);

```

2.8 Applications of Frequency Domain Analysis

Frequency domain analysis is a technique for decomposing complex signals into simpler signals. Many physical signals can be expressed as the sum of simple signals

Fig. 2.59 Analyzing window functions using MATLAB. **a** time domain signal; **b** signal spectrum; **c** window function in time domain; **d** windowed signal in time domain; **e** spectrum of windowed signal



of different frequencies. The process of finding out the information (may be magnitude, power, intensity or phase, etc.) at different frequencies is called spectrum analysis, which is mainly used to identify periodic components in the signal and is the most common method in signal analysis. Examples of its applications are as follows:

1. Spectral array analysis of automobile acceleration muffler

Automobile noise causes great harm to humans, seriously pollutes the urban environment, and affects our work, life and health. Noise control is not only related to the comfort of driving, but also related to the protection of the environment. Engine noise is the main component of automobile noise. It is a combination of noise emitted by multiple sound sources, mainly including aerodynamic noise and structural noise. Aerodynamic noise mainly includes intake noise, exhaust noise, airflow noise caused by superchargers and fans; structural noise is mainly formed by surface vibration caused by the combustion excitation force and mechanical excitation force of the structure. Among engine noise sources, exhaust noise accounts for about 30% of the total noise, which is the largest proportion of all noise sources, and is 10–15 decibels higher than other machine noises. There are many types of car mufflers, and their principles are also different. They can be mainly divided into resistive mufflers, reactive mufflers, impedance composite mufflers, micro-perforated plate mufflers, small hole mufflers and active mufflers. The spectral array analysis of the muffler is helpful for people to design, select and optimize the structure of the muffler according to actual application scenarios (Fig. 2.60).

2. Vibration measurement of CNC machine tools

Computer numerical controlled (CNC) machine tools are the key equipment of modern manufacturing. In order to obtain a numerically controlled machine tool with high machining accuracy and processing efficiency, the vibration of the machine tool must be considered in the research and development process. Especially for high-end CNC machine tools with a maximum spindle speed of 2000–40,000 r/min,

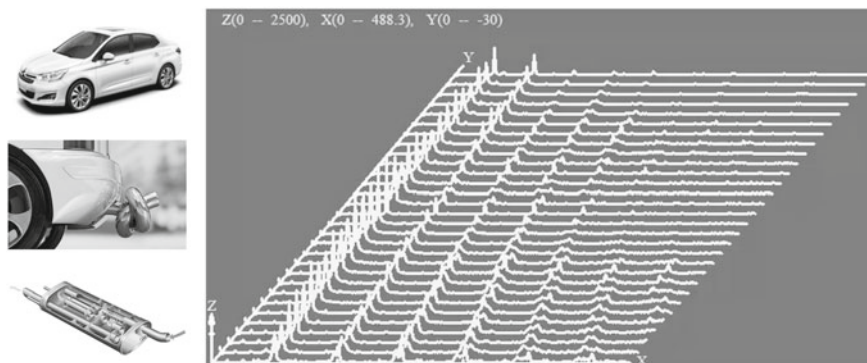


Fig. 2.60 Spectral array analysis of car muffler

a maximum feed speed of 120 m/min, and a machining accuracy in the order of micrometers or even nanometers, the vibration analysis and vibration suppression are particularly important. The vibration measurement of the CNC machine tool can be used to evaluate the anti-vibration performance of the machine tool and verify the correctness of the vibration analysis. It is an important way to solve the vibration problem of the CNC machine tool. The measurement content mainly includes natural frequency, damping ratio, mode shape, dynamic stiffness and vibration response characteristics (Fig. 2.61).

3. Bridge vibration frequency analysis

With the development of society, number of vehicles increases rapidly and the problem of traffic jam is gradually becoming serious. Bridges have been overloaded for a long time. In order to ensure the safety of bridges, bridge inspection is extremely important. The vibration frequency analysis of the bridge is one of the important means of bridge state assessment. Using the frequency domain signals obtained by dynamic measurement, the vibration characteristics and modal parameters (natural frequency, damping, mode shape, etc.) of bridges can be obtained. And the load-bearing capacity of bridges (operating time, limited load vehicle weight, etc.) can be evaluated.

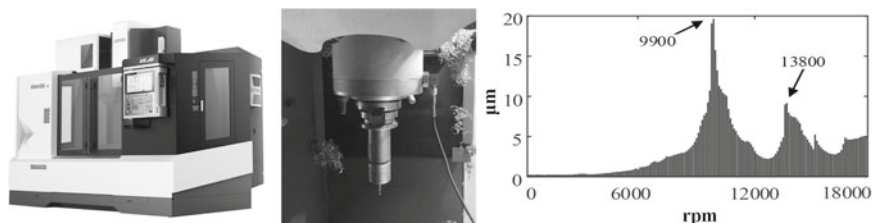


Fig. 2.61 Natural frequency of CNC machine tools

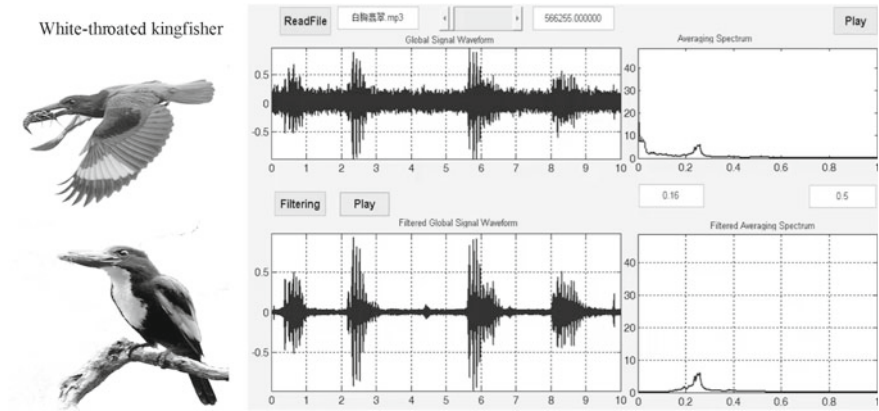


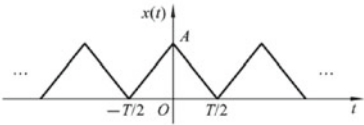
Fig. 2.62 Spectrum analysis and filtering of bird song

4. Spectrum analysis and filtering of bird call

In the research of bird ecology, it is very useful to record bird calls to assess bird richness and diversity. At present, there are many automatic recording devices and recognition software, which can distinguish bird types by monitoring their sounds and performing spectrum analysis. It effectively improves the efficiency of ecological research. Figure 2.62 shows the spectrum analysis of the sound of white-throated kingfisher.

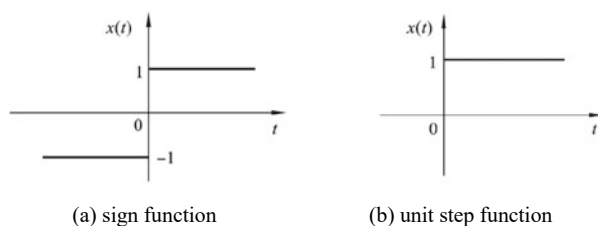
Exercise

- 2.1 Discuss the respective characteristics of the spectra of periodic signals and aperiodic signals
- 2.2 Explain the aliasing, spectral leakage and fence effects in the process of signal discretization, and explain how to prevent these phenomena from occurring.
- 2.3 Briefly explain the influence of the window function on the truncated signal spectrum.
- 2.4 A periodic triangular signal is shown in Exercise Fig. 2.1, find its Fourier series in trigonometric form and complex exponential form, and draw the spectra.



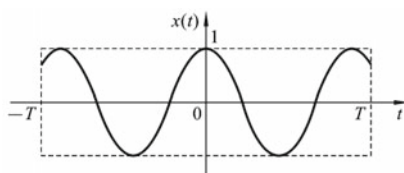
Exercise Fig. 2.1

- 2.5 Find the spectra of the sign function (Exercise Fig. 2.2a) and the unit step function (Exercise Fig. 2.2b).



Exercise Fig. 2.2

- 2.6 Find the spectrum of an exponentially decay signal $x(t) = x_0 e^{-at} \sin(\omega_0 t + \phi_0)$.
- 2.7 Find the Fourier transform of the truncated cosine function $\cos \omega_0 t$ (Exercise Fig. 2.3), and use the properties of Fourier transform to analyze the effect of truncation on the signal spectrum.



Exercise Fig. 2.3

- 2.8 Let $x[n]$ be a finite-length sequence of length N , and $\text{DFT}(x[n]) = X[k]$. Padding the data with zeros to double its length to get a signal $y[n]$ with length of $2N$. Analyze the relationship between $\text{DFT}(y[n])$ and $X[k]$, and explain the effect of zero-padding.

Chapter 3

Amplitude Domain Analysis



The signal analyzing method with the amplitude of signal as abscissa is called the amplitude domain analysis, mainly includes: Probability density curve, histogram and probability distribution curve. In the monitoring of rotating machinery operation state, the amplitude domain analysis of the vibration signal can be used to determine whether the machine is operating normally; in mechanical experiment, the fatigue damage can be estimated through the amplitude domain analysis.

3.1 Probability Density Function and Histogram

Probability density function analysis is a statistical analysis method in which the magnitude of the amplitude is taken as the abscissa and the probability of occurrence within each amplitude interval is taken as the ordinate. The probability density function of the signal is defined as:

$$p(x) = \lim_{\Delta x \rightarrow 0} \frac{P[x < x(t) \leq x + \Delta x]}{\Delta x} \quad (3.1)$$

where, $P[x < x(t) \leq x + \Delta x]$ represents the probability of the instantaneous value of $x(t)$ falling into the interval $[x, x + \Delta x]$.

As shown in Fig. 3.1, the observation interval is $[0, T]$, and the total time of the instantaneous value $x(t)$ falling into the interval $[x, x + \Delta x]$ is

$$T_x = \Delta t_1 + \Delta t_2 + \dots + \Delta t_n \quad (3.2)$$

When observation time goes to infinity $T \rightarrow \infty$, the limit of T_x/T is the probability that the instantaneous value of the signal $x(t)$ falls within the range $[x, x + \Delta x]$, namely:

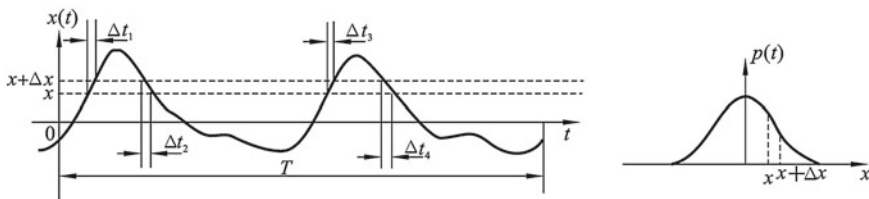


Fig. 3.1 Calculation of signal probability density function

$$P[x < x(t) \leq x + \Delta x] = \lim_{T \rightarrow \infty} \frac{T_x}{T} \quad (3.3)$$

And the probability density

$$p(x) = \lim_{T \rightarrow \infty, \Delta x \rightarrow 0} \frac{T_x}{T} \left[\left(\frac{T_x}{T} \right) / \Delta x \right] \quad (3.4)$$

represents the probability that the instantaneous value falls within the incremental range Δx .

Figure 3.1 shows the process of calculating probability density according to Eq. (3.4). The obtained probability density function $p(x)$ is a function of the amplitude x of signal $x(t)$, which reflects the probability of the signal possessing different amplitudes. It is often used in engineering to determine the signal properties and determine whether the measurement signal is normal. The probability density function can be used to express the mean value, the mean square value or the variance, according to probability theory.

The first-order moment of origin, i.e. the mean:

$$\mu_x = \int_{-\infty}^{\infty} x p(x) dx \quad (3.5)$$

The second-order moment of origin, i.e. the mean square value:

$$\psi_x^2 = \int_{-\infty}^{\infty} x^2 p(x) dx \quad (3.6)$$

The third-order moment of origin, i.e. the variance:

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 p(x) dx \quad (3.7)$$

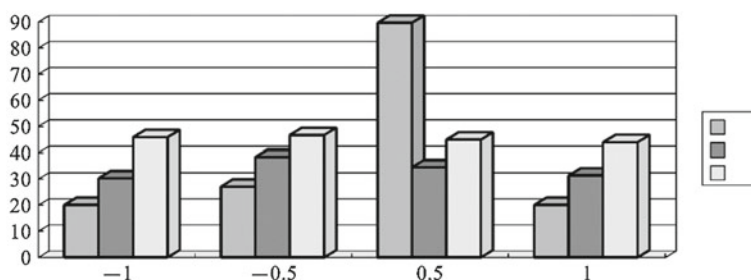


Fig. 3.2 Histogram

Histogram is a representation of the distribution of discrete data. To get the histogram, the whole amplitude range should be divided into a series of intervals, then count the number of times of signal values falling into each interval. A histogram is plot with the amplitude interval as abscissa and the number of occurring times as ordinate, as shown in Fig. 3.2. The probability density function is obtained by normalizing the histogram and let the interval become infinitesimal.

Example 3.1 The Waveforms and Probability Density Functions of Typical Signals

The probability density functions of some typical signals, including square wave, sinusoidal wave, sinusoidal wave plus random signal, narrowband random signal and wideband random signal, are shown in Fig. 3.3.

Example 3.2 Use Probability Density Function to Diagnose Bearing Fault

Figure 3.4a shows the vibration signal of a healthy rolling bearing vibration signal waveform and its probability density curve, which is a typical normal distribution curve; Fig. 3.4b shows the signal of a bearing with defects and its probability density curve. Since defects cause impacts to the bearing, the probability density curve will be skewed and scattered.

Example 3.3 Use MATLAB to get the histogram and probability density function of the sinusoidal signal

The following MATLAB code shows the calculation of histogram and probability density function of a sine function of 50 Hz. The result is shown in Fig. 3.5.

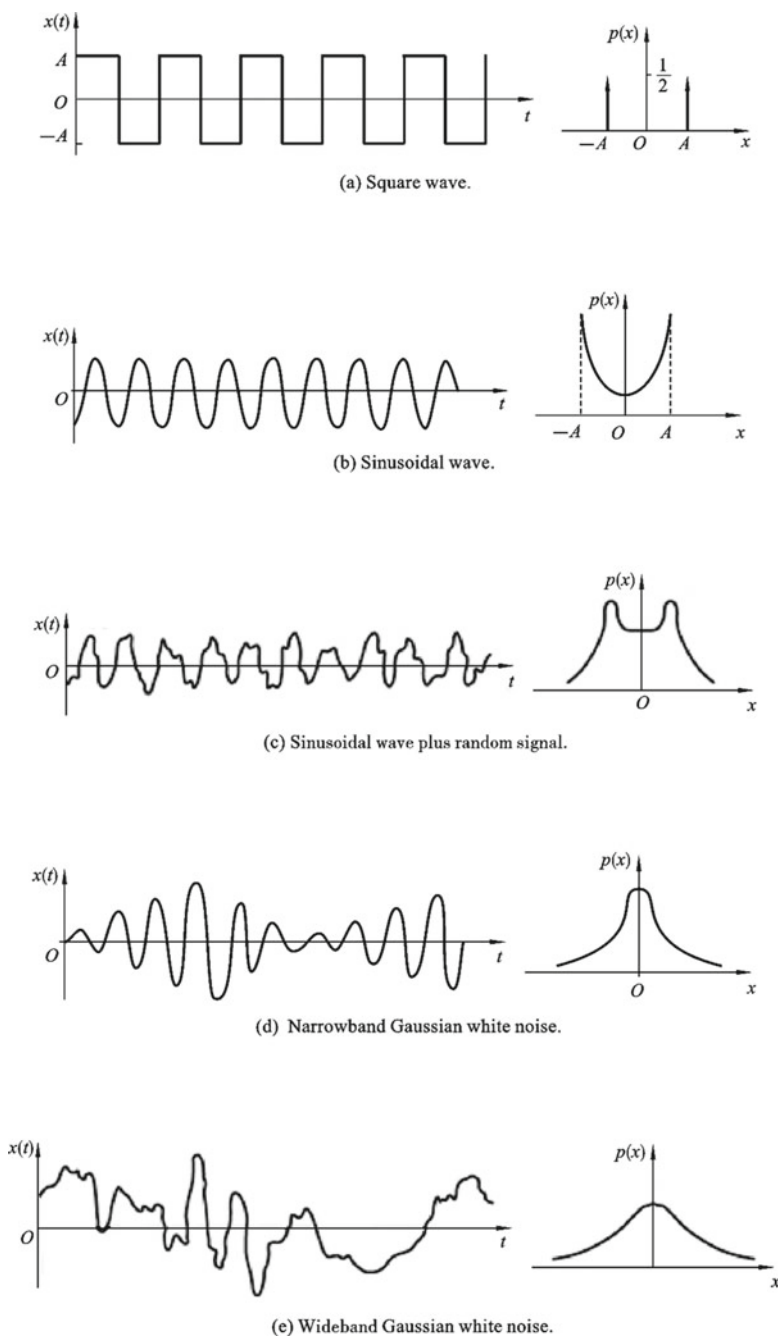


Fig. 3.3 Waveforms and probability density functions of several typical signals

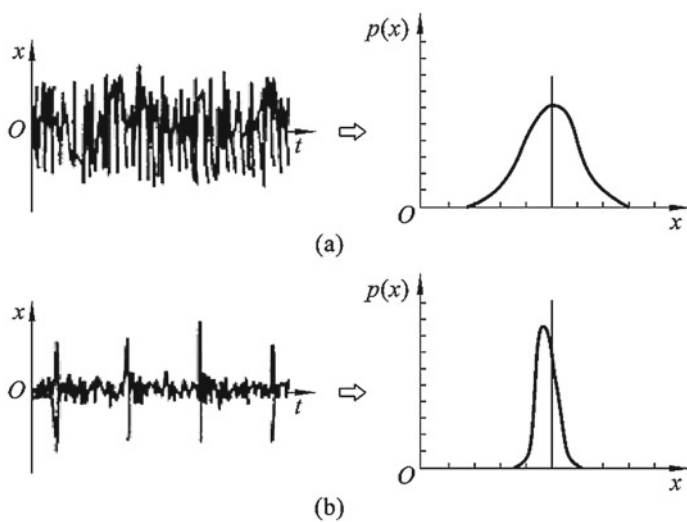


Fig. 3.4 Rolling bearing fault diagnose by probability density

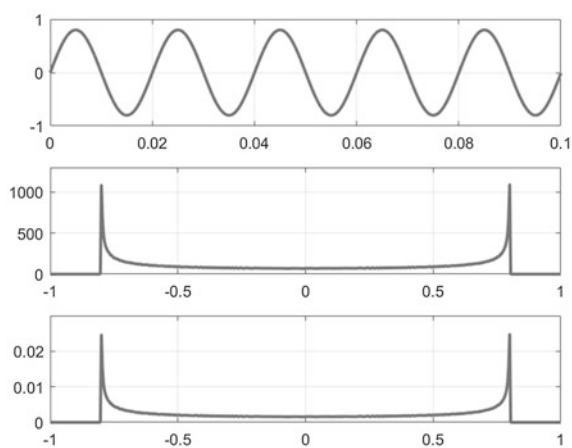


Fig. 3.5 Calculate the histogram and probability density function of the sinusoidal signal using MATLAB

```

Fs=44100; dt=1.0/Fs; T=1; N=T/dt;
x=linspace(0,T,N);
y=0.8*sin(2*3.14*50*x);
subplot(3,1,1); plot(x,y,'linewidth',2);
xlim([0,0.1]); ylim([-1,1]); grid on;
A1=-1; A2=1; M=500; da=(A2-A1)/M;
a=linspace(A1,A2,M);
h=hist(y,a);
subplot(3,1,2); plot(a,h,'linewidth',2);
ylim([0,1300]); grid on;
p=h/length(y);
subplot(3,1,3); plot(a,p,'linewidth',2);
ylim([0,0.03]); grid on;

```

Example 3.4 Use MATLAB to calculate the probability density function of the sawtooth wave

The following MATLAB code shows the calculation of probability density function of a sawtooth wave of 50 Hz. The result is shown in Fig. 3.6.

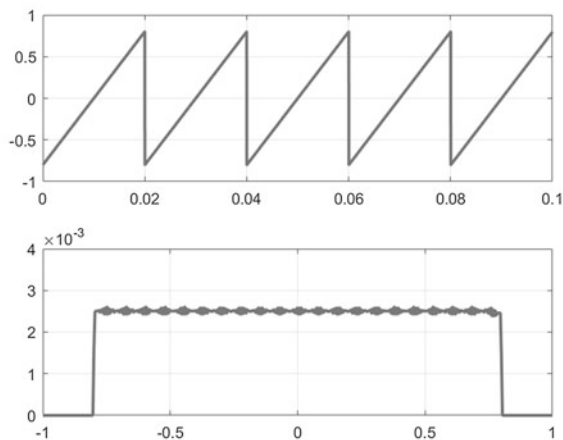


Fig. 3.6 Calculating the probability density function of the sawtooth wave using MATLAB

```

Fs=44100; dt=1.0/Fs; T=1; N=T/dt;

x=linspace(0,T,N);

y=0.8*sawtooth(2*3.14*50*x);

subplot(2,1,1); plot(x,y,'linewidth',2);

xlim([0,0.1]); ylim([-1,1]);

grid on;

A1=-1; A2=1; M=500;

da=(A2-A1)/M;

a=linspace(A1,A2,M);

h=hist(y,a);

p=h/length(y);

subplot(2,1,2); plot(a,p,'linewidth',2);

ylim([0,0.004]); grid on;

```

3.2 Probability Distribution Function

Cumulative distribution function, also known as probability distribution function, is one of the basic concepts of probability theory. In engineering practice, it is often necessary to study the probability of a random variable ξ taking a value less than a certain value x . The probability value is a function of x , and is called the distribution function of random variable ξ , denoted as $F(x) = P(\xi < x)$ ($-\infty < x < \infty$). It can calculate the probability that a random variable falls within any range. For example, in the design of bridges and dams, the probability that the highest water level ξ of the river is less than x meters per year is a function of x , which is the distribution function of the highest water level ξ . The commonly used distribution functions include normal distribution function, Poisson distribution function and binomial distribution function.

The cumulative distribution function is a mathematical expression that describes the distribution of random variables. For any real number x , the probability of the event $[X < x]$ is of course a function of x . Assume

$$F(x) = P(X < x) \quad (3.8)$$

Obviously, we have

$$\begin{cases} F(-\infty) = 0 \\ F(\infty) = 1 \end{cases} \quad (3.9)$$

and $F(x)$ is called the cumulative distribution function of random variable X . Thus, the probability of an event $[a \leq X \leq b]$ happening is determined by the cumulative distribution function $F(x)$. In another word, the cumulative distribution function $F(x)$ completely describes the statistical characteristics of the random variable X .

3.2.1 Cumulative Distribution of Discrete Random Variables

For a discrete random variable X , assume x_1, x_2, \dots, x_n are values of X , and p_1, p_2, \dots, p_n are the corresponding probabilities of the above values. Then the probability distribution of the discrete random variable X is:

$$P(X = x_i) = p_i, \quad i = 1, 2, \dots, n \quad (3.10)$$

and the probabilities p_i satisfy:

$$\sum_{i=1}^n p_i = 1 \quad (3.11)$$

Therefore, the cumulative distribution function of the discrete random variable X is:

$$F(x) = \sum_{x_i < x} p_i \quad (3.12)$$

3.2.2 Probability Distribution of Continuous Random Variables

For a continuous random variable X with its value falls within the interval $[a, b]$, assume its distribution function $F(x)$ is a monotonically increasing function, differentiable in the interval $(-\infty, \infty)$ and its derivative $F'(x)$ is continuous. Then, the probability of X falling into the interval $[x, x + \Delta x]$ is:

$$P(x \leq X \leq x + \Delta x) = F(x + \Delta x) - F(x) \quad (3.13)$$

In order to describe its probability distribution, the concept of probability distribution density function is introduced

$$f(x) = F'(x) = \lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} \quad (3.14)$$

Therefore, the cumulative distribution function of the continuous random variable X can be written in the form of a commonly used probability integral:

$$F(x) = \int_{-\infty}^x f(x) dx \quad (3.15)$$

In this way, as long as the probability distribution density function of a continuous random variable X is known, the probability that X falls within a certain interval can be obtained:

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x) dx \quad (3.16)$$

Similar to the probability function of discrete random variable, for the probability density function of continuous variable, we have

$$f(x) \geq 0 \quad (3.17)$$

and

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (3.18)$$

The probability density functions of continuous random variables $f(x)$ and their corresponding distribution function $F(x)$ are shown in Fig. 3.7a and b, respectively. Sometimes the graph of $f(x)$ is called the probability density curve, and the graph of $F(x)$ is called the cumulative distribution curve.

Example 3.5 For the following sinusoidal signal, find its probability density function and probability distribution function.

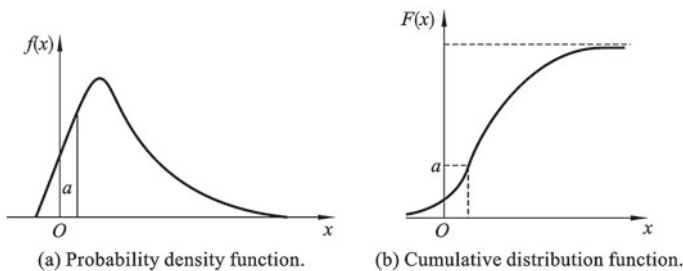
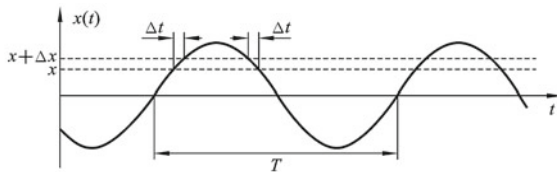


Fig. 3.7 The probability density curve and cumulative distribution curve

Fig. 3.8 Probability density calculation of sinusoidal signal



$$x(t) = A \sin(2\pi f_0 t + \varphi)$$

From the signal expression we get:

$$2\pi f_0 t + \varphi = \arcsin \frac{x}{A}$$

Take the derivative and we get:

$$\frac{dt}{dx} = \frac{1}{2\pi f_0} \frac{1}{\sqrt{A^2 - x^2}}$$

The sinusoidal signal is a periodic signal. We can take an observation time of $T = 1/f_0$ to analyze the probability density of the signal, as shown in Fig. 3.8.

As shown in Fig. 3.8, within one period, the time corresponding to signal value falling into the interval $(x, x + \Delta x)$ is $2\Delta t$. Thus the probability density is:

$$f(x) = \lim_{\Delta x \rightarrow 0} \frac{1}{\Delta x} \left[\frac{\sum \Delta t}{T} \right] = \frac{1}{T} \frac{2dt}{dx}$$

Substitute the derivate inside, the probability density can be obtained:

$$f(x) = \frac{1}{\pi \sqrt{A^2 - x^2}}$$

Integrating the probability density, we get:

$$\begin{aligned} F(x) &= \int_{-\infty}^x \frac{1}{\pi \sqrt{A^2 - x^2}} dx \\ &= \frac{1}{\pi} \arcsin\left(\frac{x}{A}\right) \Big|_{-A}^x \\ &= \frac{1}{\pi} \left[\arcsin\left(\frac{x}{A}\right) + \frac{\pi}{2} \right] \end{aligned}$$

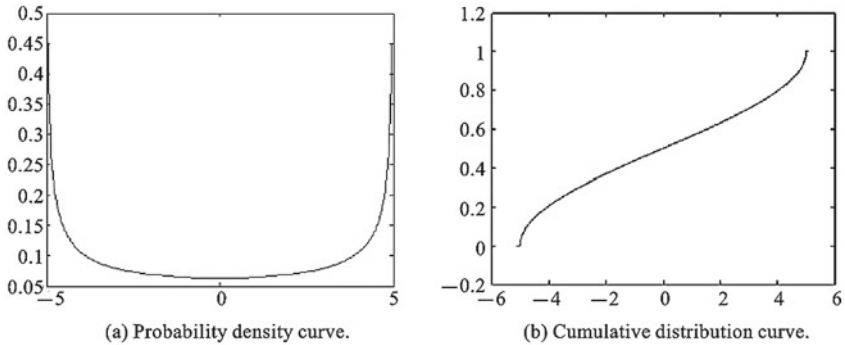


Fig. 3.9 The probability density curve and cumulative distribution curve of the sine signal

Assuming the amplitude of the sine wave signal is $A = 5$, the probability density curve and probability distribution curve of the signal is shown in Fig. 3.9.

Example 3.6 Calculating the cumulative distribution function of a sinusoidal signal using MATLAB.

The following MATLAB code is to show the process of calculating the probability distribution function of a sinusoidal signal of 50 Hz. The result is shown in Fig. 3.10.

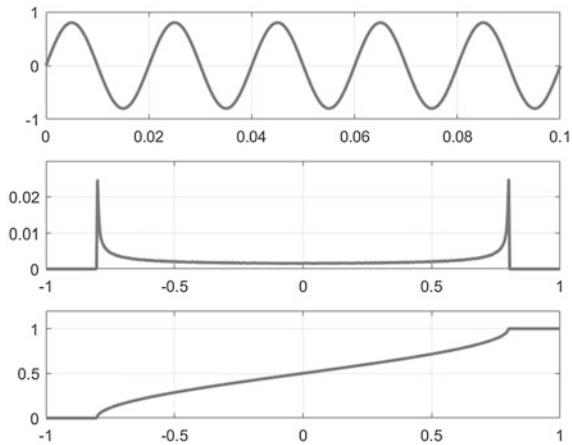


Fig. 3.10 Calculated probability distribution function of sine signal

```

Fs=44100; dt=1.0/Fs; T=1; N=T/dt;

x=linspace(0,T,N);

y=0.8*sin(2*3.14*50*x);

subplot(3,1,1); plot(x,y,'linewidth',2);

```

```

xlim([0,0.1]); ylim([-1,1]); grid on;

A1=-1; A2=1; M=500; da=(A2-A1)/M;

a=linspace(A1,A2,M);

h=hist(y,a);

p=h/length(y);

subplot(3,1,2); plot(a,p,'linewidth',2);

ylim([0,0.03]); grid on;

c(1)=0;

for i=2:M

    c(i)=c(i-1)+p(i);

end

subplot(3,1,3); plot(a,c,'linewidth',2);

xlim([-1,1]);ylim([0,1.2]); grid on;

```

3.3 Engineering Applications of Amplitude Domain Analysis

3.3.1 Machine Fault Diagnosis

Machine fault diagnosis is a technology that monitors the operating state of machine. Vibration monitoring, noise monitoring, oil monitoring, performance trend analysis and non-destructive testing are the main diagnostic techniques. Diagnostic technology has become a popular research topic and brings a lot of economic benefits.

Example 3.7 Diagnosis of Air Conditioning Noise

When the air conditioning is running, noise is generated. The noise waveform is shown in Fig. 3.11. By analyzing the probability density distribution of the air conditioning noise, it is easy to diagnose the healthy status of the air conditioning and find the abnormal state such as loose installation or buzzing.

Example 3.8 Fault Diagnosis of the Gearbox

The time-domain waveforms of normal gearbox vibration and its probability density are shown in Fig. 3.12a. The waveforms and probability corresponding to the gearbox with crack and wearing are shown in Fig. 3.12b and c. By analyzing the probability density distribution, the different states of the transmission gear can be easily diagnosed.

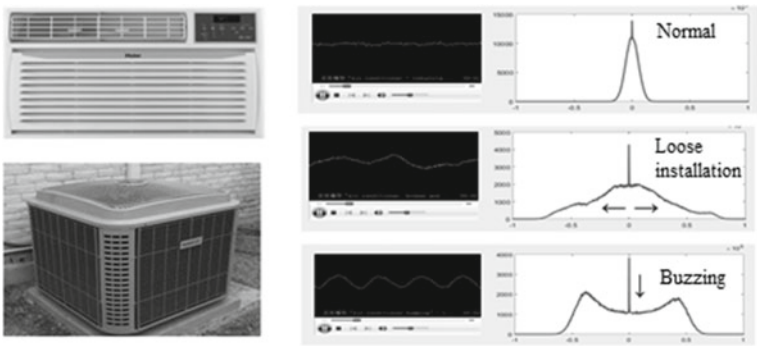


Fig. 3.11 Air conditioning noise diagnosis

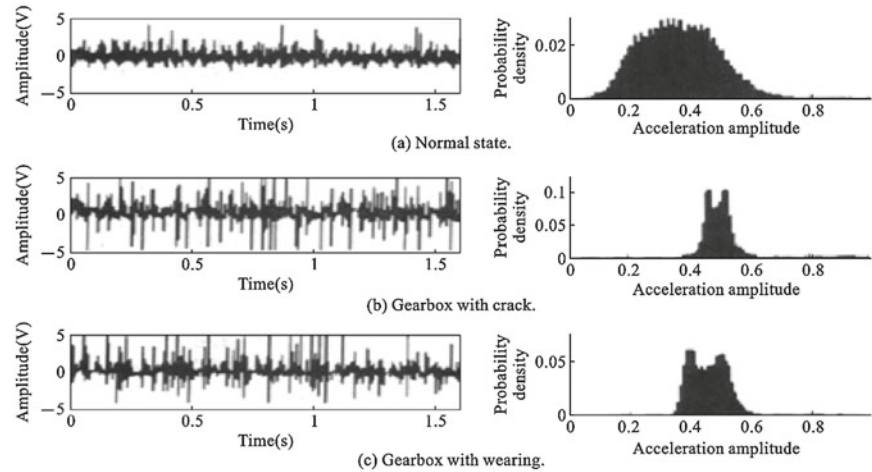


Fig. 3.12 Gearbox fault diagnosis

3.3.2 *Histogram Analysis of Photo Quality*

Histogram analysis is a very useful tool. Digital cameras provide image histogram analysis, which can help users quickly obtain image quality reports, as shown in Fig. 3.13.

A digital image consists of a 2-D array of pixels. Each pixel in the array can be decomposed into red, green, and blue (RGB). The brightness value of each color is between 0 and 255. The histogram is the brightness distribution density of all pixels in each color or grayscale image, as shown in Fig. 3.14.

Histogram equalization is an image enhancement method. Its goal is to map the unevenly distributed colors in the original image to distribute them uniformly to improve the contrast of the image.



Fig. 3.13 Photos taken with a digital camera

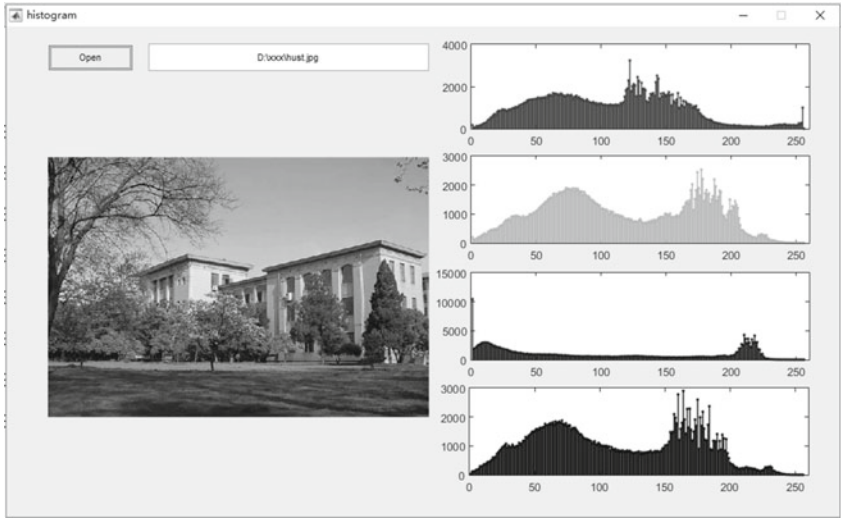


Fig. 3.14 Histogram analysis of photo quality

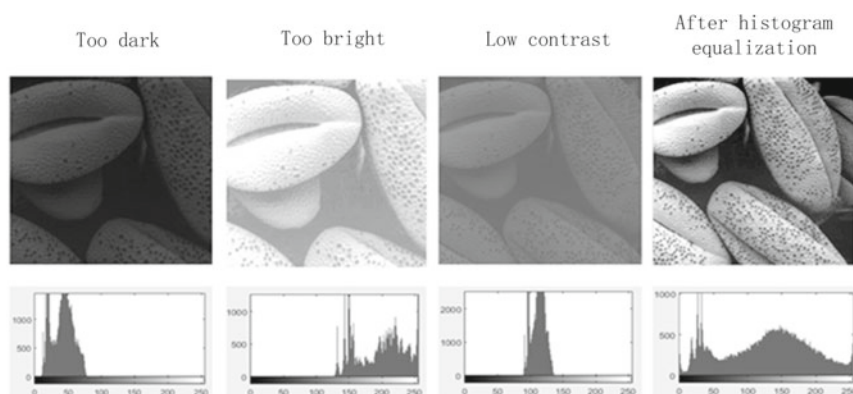


Fig. 3.15 Histogram equalization of grayscale image

Example 3.9 Histogram Equalization of Grayscale Images

It can be seen from Figs. 3.3, 3.15 that through the histogram equalization of the grayscale image, the image quality can be significantly improved for the image that is too dark, too bright, or low contrast.

Example 3.10 Histogram Equalization of Color Images

It should be noted that the histogram equalization cannot be applied directly to the RGB channels. The RGB image needs to be converted to an HSV image (hue H, saturation S, brightness V), then only the V channel is equalized.

Exercise

- 3-1 Briefly describe the definitions and differences of probability density function, histogram and probability distribution function.
- 3-2 Find the probability density function and probability distribution function of the signal $x(t) = A \sin(\omega_0 t + \varphi)$.
- 3-3 Design a GUI program interface with MATLAB to draw the probability density curve and probability distribution curve of four signals: Sine wave, square wave, triangular wave and white noise.
- 3-4 List several applications of amplitude domain analysis in engineering, and briefly describe their principles.
- 3-5 Briefly describe the definition, principle and application of histogram equalization.

Chapter 4

Correlation Analysis of Signals



In signal analysis, sometimes it is necessary to study the relationship between two or more signals, such as in communication systems, radar systems, and control systems. In these examples, the signal waveform at the emitting end is known. In the receiving end, we must find whether the emitted signal is received. Even if the signal sent by the emitting end is presented in the received signal, it is difficult to identify it due to signal distortion caused by various interferences. An intuitive idea is to compare the emitted signal with the received signal, and use their degree of similarity to make a judgment. Thus, we need to solve the problem of measuring the similarity of signals, which brings to the topic of this chapter, namely the correlation analysis of signals in the time-different domain. Correlation analysis is a commonly used method to determine the degree of similarity between two signals in the processing of measurement signals. It is widely used in engineering applications such as signal separation, signal delay estimation, and target positioning.

4.1 Concept of Correlation Analysis

4.1.1 Correlation of Variables

In statistics, the correlation coefficient is used to describe the correlation between variables x and y . It is the mathematical expectation of the product of two random variables and represents the degree of correlation between x and y , and it is defined as:

$$\rho_{xy} = \frac{c_{xy}}{\sigma_x \sigma_y} = \frac{\sum [(x_i - \mu_x)(y_i - \mu_y)]}{\sqrt{\sum (x_i - \mu_x)^2 \times \sum (y_i - \mu_y)^2}} \quad (4.1)$$

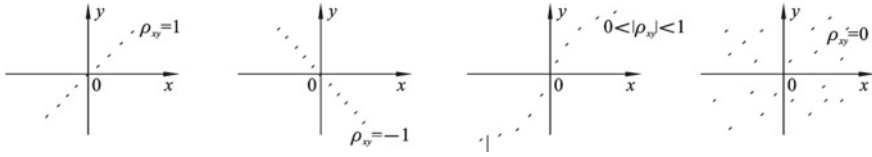


Fig. 4.1 Different types of correlation between variables x and y

where, c_{xy} is the mathematical expectation of the product of the fluctuations of two random variables, σ_x and σ_y are mean square errors of variables x and y , μ_x and μ_y are mean values of x and y .

It can be proven that the correlation coefficient satisfies $-1 \leq \rho_{xy} \leq 1$. When x and y are completely correlated, $\rho_{xy} = \pm 1$; when x and y are completely unrelated, $\rho_{xy} = 0$; when x and y are partially correlated, $0 < |\rho_{xy}| < 1$. Different types of correlation between variables x and y are shown in Fig. 4.1.

Many things in nature are related, such as our height and weight, amount of smoking and life span, stress and strain in structures, etc. We can use the correlation coefficient to judge the relationship between them through statistical calculations.

4.1.2 Correlation of Signals

If the studied variables x and y are functions of time, i.e. $x(t)$ and $y(t)$. A time-related quantity τ can be introduced into the correlation coefficient, then it becomes the correlation coefficient function. It reflects the degree of similarity and correlation between two signals at different time differences, i.e. the correlation in time-shifting coordinates. Therefore, correlation analysis of time domain signals is also called time-difference domain analysis or time delay domain analysis. When $x(t)$ and $y(t)$ are energy signals, the correlation coefficient function is:

$$\rho_{xy}(\tau) = \frac{\int_{-\infty}^{\infty} x(t)y(t+\tau)dt}{[\int_{-\infty}^{\infty} x^2(t)dt \int_{-\infty}^{\infty} y^2(t)dt]^{1/2}} \quad (4.2)$$

Similarly, we can prove that $-1 \leq \rho_{xy}(\tau) \leq 1$. The correlation function $\rho_{xy}(\tau)$ is a function of time difference τ , and when $x(t)$ and $y(t)$ are completely correlated after a time shifting of τ , $\rho_{xy}(\tau) = \pm 1$; when $x(t)$ and $y(t)$ are completely unrelated after a time shifting of τ , $\rho_{xy}(\tau) = 0$; when $x(t)$ and $y(t)$ are partially correlated after a time shifting of τ , $0 < |\rho_{xy}(\tau)| < 1$.

4.1.3 Cross-Correlation Function

For the sake of simplicity, sometimes the correlation coefficient function is not normalized in engineering. The cross-correlation function is defined as:

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)y(t + \tau)dt \quad (4.3)$$

or

$$R_{yx}(\tau) = \int_{-\infty}^{\infty} y(t)x(t + \tau)dt \quad (4.4)$$

$R_{xy}(\tau)$ and $R_{yx}(\tau)$ are called the cross-correlation of the two signals.

4.1.4 Auto-Correlation Function

If $y(t) = x(t)$, then $R_{xy}(\tau)$ becomes $R_{xx}(\tau)$, which is called the auto-correlation and usually abbreviated as $R_x(\tau)$. It is mathematically expressed as:

$$R_x(\tau) = \int_{-\infty}^{\infty} x(t)x(t + \tau)dt \quad (4.5)$$

If $x(t)$ and $y(t)$ are power signals, then the correlation functions are defined as:

$$R_{xy}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)y(t + \tau)dt \quad (4.6)$$

$$R_{yx}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} y(t)x(t + \tau)dt \quad (4.7)$$

$$R_x(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t + \tau)dt \quad (4.8)$$

It can be seen that the dimensions of correlation function of energy signal and power signal are different, the former is energy, and the latter is power.

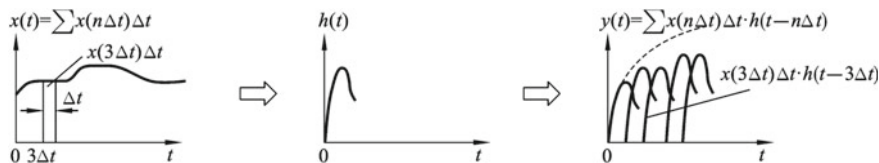


Fig. 4.2 Describe the system response using convolution integral

4.1.5 Convolution

Convolution is a mathematical method that characterizes the input–output relationship of a linear time-invariant system. It plays an important role in the study of system characteristics. The convolution integral of the function $x(t)$ and $h(t)$ is defined as:

$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(t - \tau)d\tau = x(t) * h(t) \quad (4.9)$$

In the analysis of linear time-invariant systems, convolution integral has a clear physical meaning. Suppose the unit impulse response function of the system is $h(t)$, and the signal $x(t)$ is decomposed into impulse signals. According to the superposition of the linear system, the response of the system can be decomposed into the sum of the impulse signal responses, as shown in Fig. 4.2.

If Δt approaches infinitesimal, then the summation of discrete sequence in Fig. 4.2 becomes an integral, and the convolution integral equation in Eq. (4.9) is obtained.

In order to deepen the understanding of convolution integral, the process of convolution integral is graphically illustrated in Fig. 4.3. Figure 4.3a is the waveform of the signal $x(\tau)$, Fig. 4.3b is the waveform of the signal $h(\tau)$; according to the equation of convolution integral, the signal $h(\tau)$ is first flipped to obtain the waveform of $h(-\tau)$ (Fig. 4.3c); then shift $h(-\tau)$ to obtain $h(t - \tau)$ (Fig. 4.3d); then multiply $h(t - \tau)$ by $x(\tau)$ to get the new waveform in Fig. 4.3e; integrate $x(\tau) \bullet h(t - \tau)$ to get the result of convolution integral at time shift of t (Fig. 4.3f); finally, change the shifting value to get the convolution integral at different shifts, and connect them to get the convolution integral curve (Fig. 4.3g).

4.1.6 Convolution Theorem

If the Fourier transform of time domain signal $x(t)$ and $y(t)$ are respectively $X(f)$ and $Y(f)$, i.e. $x(t) \leftrightarrow X(f)$ and $y(t) \leftrightarrow Y(f)$, where the sign \leftrightarrow denotes the Fourier transform pairs, then

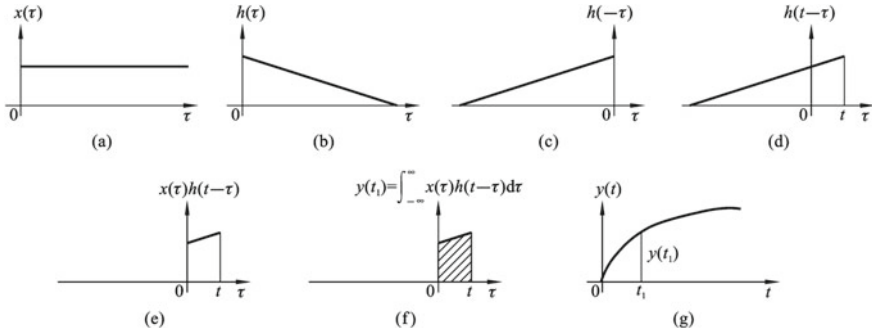


Fig. 4.3 Graphical illustration of the convolution integral process

$$x(t) * y(t) \leftrightarrow X(f)Y(f)$$

$$x(t)y(t) \leftrightarrow X(f) * Y(f)$$

The equations indicate that the convolution in time domain corresponds to the multiplication in frequency domain and vice versa. They are called the time-domain convolution theorem and the frequency-domain convolution theorem, respectively. For the derivation and proof of the theorems, please refer to books on Fourier transform.

The direct calculation of convolution integral is complicated. With the help of the time domain convolution theorem, we can turn the convolution integral operation in time domain into a multiplication operation in the frequency domain, and then obtain the convolution result through the inverse Fourier transform. The convolution theorem thus provides an effective way of calculating convolution.

4.2 Properties of the Correlation Function

4.2.1 Properties of Auto-Correlation Function

According to the definition of auto-correlation function

$$R_x(\tau) = \int_{-\infty}^{\infty} x(t)x(t + \tau)dt$$

it has the following properties:

- (1) The auto-correlation function is an even function of τ , as shown in Figs. 4.4 and 4.5. It satisfies the following equation:

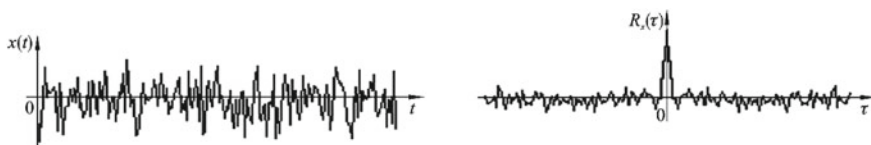


Fig. 4.4 Random signal and its auto-correlation function waveform

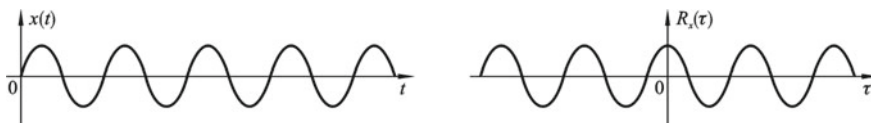


Fig. 4.5 Sine wave signal and its autocorrelation function waveform

$$R_x(-\tau) = R_x(\tau) \quad (4.10)$$

- (2) When $\tau = 0$, the auto-correlation function of $x(t)$ has the maximum value, as shown in Figs. 4.4 and 4.5. If $x(t)$ is an energy signal, then

$$R_x(0) = \int_{-\infty}^{\infty} x^2(t) dt \quad (4.11)$$

If $x(t)$ is a power signal, then

$$R_x(0) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t) dt \quad (4.12)$$

Obviously, at $\tau = 0$, the auto-correlation function of the power signal is equal to the average power of the signal, which is its mean square value. Moreover, if the mean value is zero, then the average power, auto-correlation function, and variance of the signal are all equal.

- (3) The autocorrelation function of the random noise signal decays rapidly and tends to zero as $|\tau|$ increases, as shown in Fig. 4.4.
- (4) The autocorrelation function of periodic signal is still periodic signal of the same frequency, but the phase information of the original signal is not retained, as shown in Fig. 4.5. For example, the autocorrelation function of a sinusoidal signal $x(t) = A \sin(\omega t + \varphi)$ is $R_x(\tau) = (A^2 \cos \omega \tau)/2$.

4.2.2 Properties of Cross-Correlation Function

According to the definition of cross-correlation function

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)y(t + \tau)dt$$

it has the following properties:

- (1) The cross-correlation function is neither odd nor even, as shown in Fig. 4.6.
- (2) The cross-correlation function of the two periodic signals with the same frequency is still a periodic signal with the same frequency, and the phase information of the original signal is retained, as shown in Fig. 4.6.
- (3) The cross-correlation function of the two periodic signals with different frequencies is zero, which can be proven by the orthogonality of the sine/cosine function, as shown in Fig. 4.7.

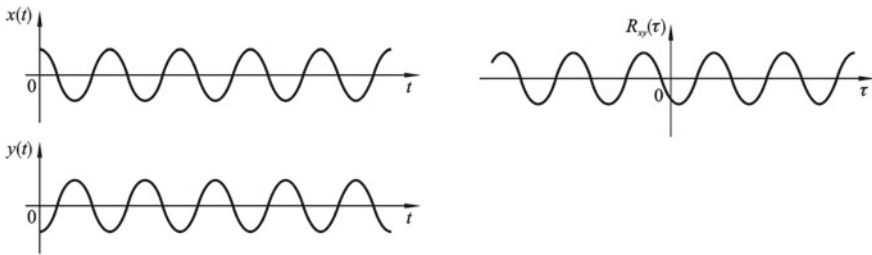


Fig. 4.6 Cross-correlation function of two signals with the same frequency

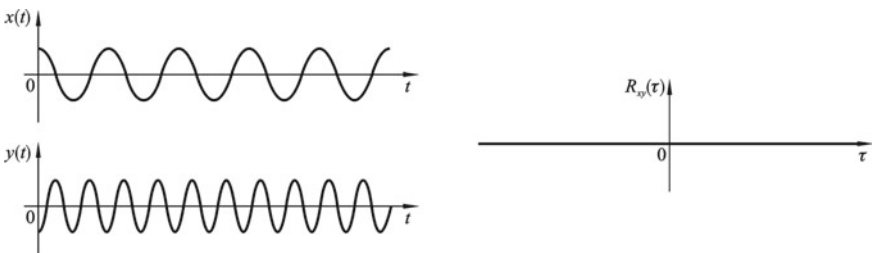


Fig. 4.7 Cross-correlation function of two signals with different frequency

4.2.3 Convolution, Correlation and Fourier Transform

Convolution is a mathematical operation used to express the relationship between the input and output of a system:

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau)h(t - \tau)d\tau \quad (4.13)$$

According to the definition of convolution integral, the convolution integral of $x(t)$ and $y(t)$ is

$$x(t) * y(t) = \int_{-\infty}^{\infty} x(\tau)y(t - \tau)d\tau \quad (4.14)$$

Correlation is a measure of the similarity between two signals. When $x(t)$ and $y(t)$ are both real energy signals, the cross-correlation function can be written as

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)y(t - \tau)dt \quad (4.15)$$

In order to facilitate the comparison, let us switch the variable t and τ in the above equation. Then the cross-correlation function can be expressed as

$$R_{xy}(t) = \int_{-\infty}^{\infty} x(\tau)y(\tau - t)d\tau = x(t) * y(-t) \quad (4.16)$$

Equations (4.14) and (4.16) show that the cross-correlation function of $x(t)$ and $y(t)$ equals the convolution of $x(t)$ and $y(-t)$.

Comparing Eqs. (4.14) and (4.16), we can find that they are closely related. As shown in Fig. 4.8, the left-hand side of the figure shows the convolution process, and the right-hand side shows the correlation process. Obviously, these two operations both include three steps of shifting, multiplication, and integration. The difference is that flipping of $y(t)$ is not required in the correlation operation, while it is required in the convolution. However, if $x(t)$ or $y(t)$ is a real even function, convolution and correlation are exactly the same.

When discussing convolution integration, we introduced that the direct calculation of convolution integration is complicated, thus we can simplify the calculation by using the time-domain convolution theorem to transform the time-domain convolution integration into a frequency-domain multiplication operation, and then apply

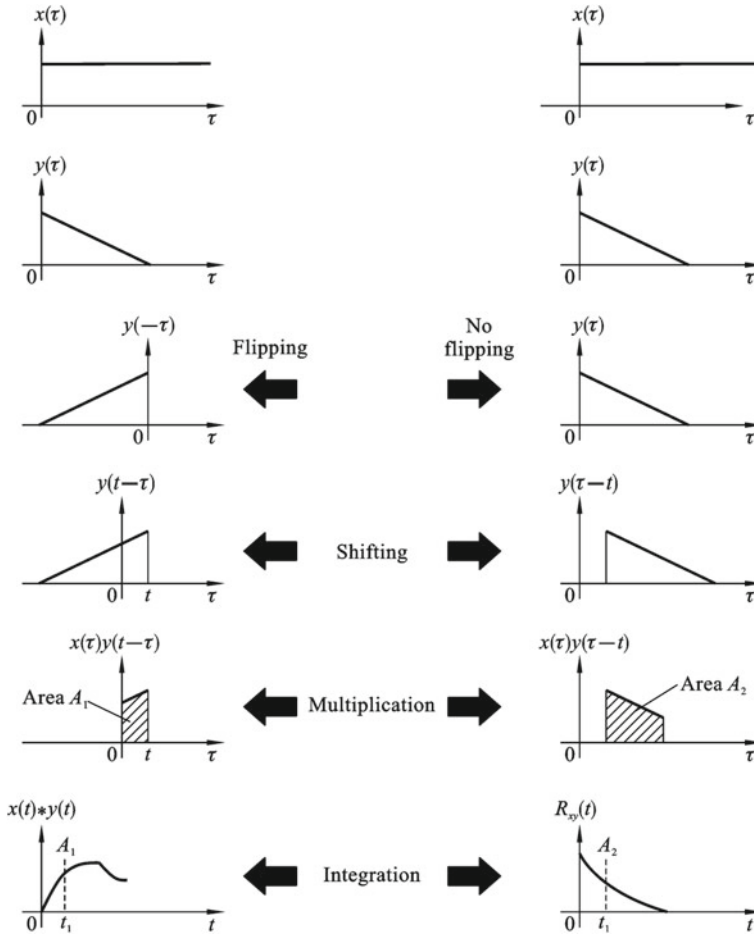


Fig. 4.8 Comparison of convolution and correlation

inverse Fourier transform to obtain the calculation result. Similarly, for correlation calculations, Fourier transforms can also be used for calculations.

The Fourier transform of convolution is the product of two Fourier transforms

$$y(t) = x(t) * h(t) \leftrightarrow Y(f) = X(f)H(f) \quad (4.17)$$

then

$$y(t) = F^{-1}[X(f)H(f)] \quad (4.18)$$

The Fourier transform of correlation is the Fourier transform of one function multiplies the conjugation of the Fourier transform of another function:

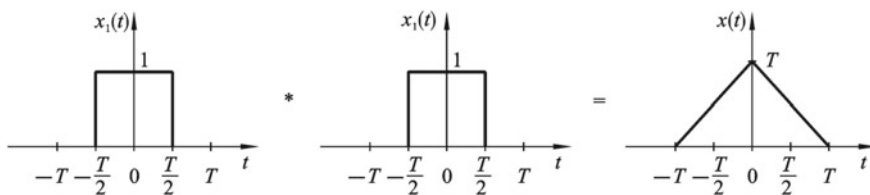


Fig. 4.9 Convolution integral of triangular pulse signal

$$R_{xy}(\tau) = \text{corr}(x, y) = \int_{-\infty}^{\infty} x(\tau)y(\tau - t)d\tau \leftrightarrow P(f) = X(f) \cdot Y^*(f) \quad (4.19)$$

Equation (4.19) is called the correlation theorem, where $Y^*(f)$ is the conjugation of $Y(f)$. Then

$$y(t) = F^{-1}[X(f)Y^*(f)] \quad (4.20)$$

Example 4.1 Find the Fourier integral of the following triangular pulse signal

$$x(t) = \begin{cases} T(1 - \frac{|t|}{T}), & |t| < T \\ 0, & \text{otherwise} \end{cases}$$

The triangular pulse signal can be regarded as the convolution integral of two rectangular pulse signals, as shown in Fig. 4.9.

The Fourier integral of a single rectangular pulse signal is

$$X_1(f) = T \frac{\sin(\pi f T)}{\pi f T}$$

According to the time domain convolution theorem, we get

$$X(f) = X_1(f)X_1(f) = [T \frac{\sin(\pi f T)}{\pi f T}]^2$$

4.3 Calculation of Correlation

4.3.1 Numerical Calculation

According to the previous definition, the cross-correlation is defined as:

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)y(t + \tau)dt$$

It can be discretized as:

$$R_{xy}[k] = \sum_0^{N-k} x[n]y[k + n] \quad k = 0, \pm 1, \pm 2 \dots$$

The above equation can be used to calculate the correlation.

4.3.2 FFT Method

The FFT method of correlation calculation is described in details in Sect. 2.6.5. This section only provides a brief supplementary explanation. The Fourier transform of correlation is equal to the Fourier transform of one function multiplies the conjugation of the Fourier transform of another function:

$$R_{xy}(\tau) = \text{corr}(x, y) \leftrightarrow P(f) = X(f)Y^*(f)$$

So, we can have an algorithm based on Fourier transform. The Fourier transform of the two signals are calculated:

$$X(f) = F(x(t))Y(f) = F(y(t))$$

Then the multiplication is taken:

$$P(f) = X(f)Y^*(f)$$

Finally, the correlation result is obtained by inverse Fourier transform.

$$R_{xy}(\tau) = F^{-1}(P(f))$$

In MATLAB, the above equation can be written as:

```
Rxy=ifft(fft(x).*conj(fft(y)))
```

It should be noted that there are also wraparound errors caused by period extension in the FFT algorithm, which can be avoided by padding zeros to the end of $x(t)$ and $y(t)$. This function has been integrated to the API function “xcorr” in MATLAB, which can be directly called by the users. However, the zero padding will cause the attenuation at both ends of the autocorrelation curve, which does not match the actual situation, as shown in Example [4.2](#)

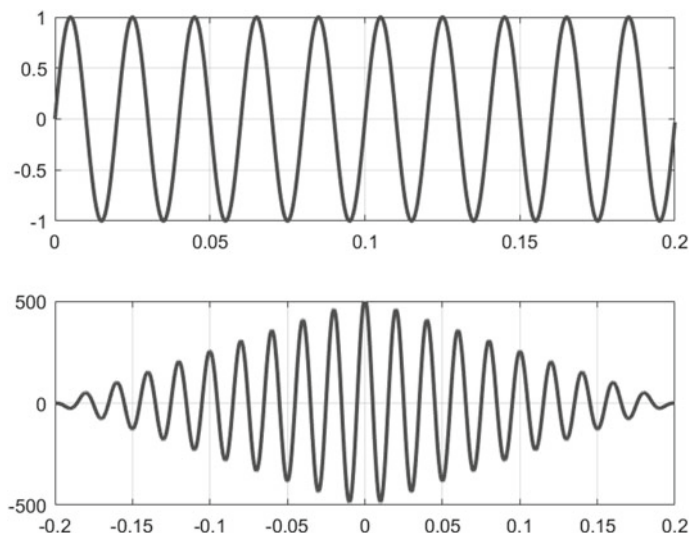


Fig. 4.10 Distortion of autocorrelation function due to zero padding

Example 4.2 Simulate the distortion of autocorrelation function due to zero padding using MATLAB The following MATLAB code is to calculate correlation with the “xcorr” function. As we can see, the calculated autocorrelation result attenuates at both ends in Fig. 4.10.

```

Fs=5120; N=1024;
dt=1.0/Fs; T=dt*N;
t=linspace(0,T,N);
x=sin(2*3.14*50*t);
subplot(2,1,1);
plot(t,x,'linewidth',2);
grid on;
r=xcorr(x);
N1=length(r);
t1=linspace(-T,T,N1);
subplot(2,1,2);
plot(t1,r,'linewidth',2);
ylim([-500, 500]);
grid on;

```

This phenomenon is also caused by period extension, which can be corrected by multiplying each point of the correlation function by a weight:

$$R_u(\tau) = R_x(\tau)/w$$

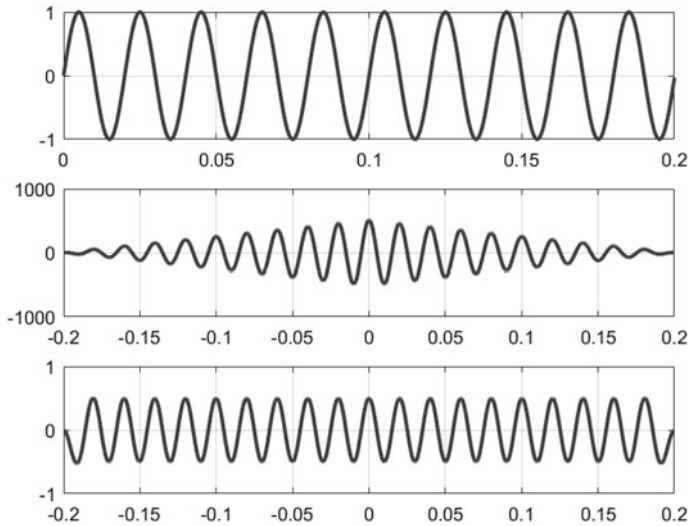


Fig. 4.11 Comparison of autocorrelation function curve before and after correction

The correction is also integrated into the API function “`xcorr`” in MATLAB. We only need to set a parameter “`unbiased`” for the “`xcorr`” function, as shown in Example 4.3.

Example 4.3 Correcting distortion caused by zero padding in MATLAB The following MATLAB code is to observe the effect of distortion correction in the calculation of autocorrelation. It is realized by setting a parameter in the `xcorr` function (Fig. 4.11).

```

Fs=5120; N=1024;
dt=1.0/Fs; T=dt*N;
t=linspace(0,T,N);
x=sin(2*3.14*50*t);
subplot(3,1,1);
plot(t,x,'linewidth',2);
grid on;
r=xcorr(x);
N1=length(r);
t1=linspace(-T,T,N1);
subplot(3,1,2);
plot(t1,r,'linewidth',2);
grid on;
r1=xcorr(x,'unbiased');
subplot(3,1,3);
plot(t1,r1,'linewidth',2);
grid on;

```

4.4 Engineering Applications of Correlation Analysis

The correlation function can be used to determine the time lag and time difference between signals. Based on this feature, it can be used in engineering for speed measurement, distance measurement and signal separation.

1. Speed measurement

Correlation analysis can be used to measure the moving speed of objects, such as the moving speed of hot-rolled steel strip, the moving speed of fluid, and the transmission speed of sound waves. Their measurement principles are similar. The measurement of sound speed is taking as an example.

In Fig. 4.12, two microphone sensors X and Y are installed along the transmission direction of the sound wave generated by the speaker S. X and Y are separated by a distance of $L = 1$ m. In the experiment, the sound wave reaches the microphone X first, and then the microphone Y. After measuring the time difference, the speed of sound can be calculated according to Eq. 4.21.

$$v = \frac{L}{\tau_0} \quad (4.21)$$

The time difference τ_0 can be determined by comparing the signals received by X and Y, and read the time difference between the similar points of their waveforms. But this method fails when there is noise interference in the measured signal. Another method is to do the cross-correlation analysis of the signal. The peak point in the cross-correlation curve corresponds to the time difference between the two signals.

2. Distance measurement

Distance measurement is actually the inverse process of speed measurement. If the moving speed of the sound is known, the distance and position of the measured object

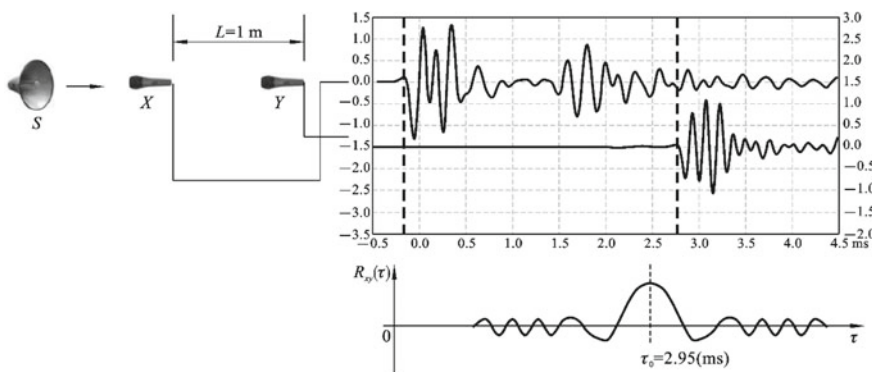


Fig. 4.12 The principle of sound speed measurement

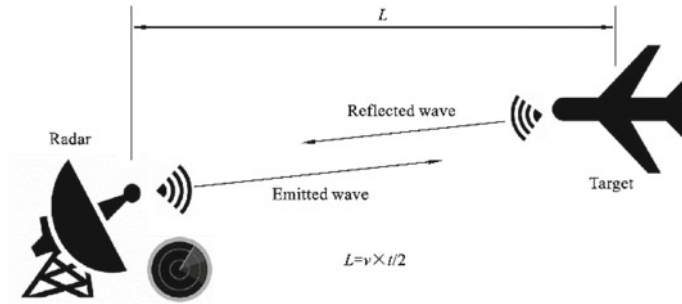


Fig. 4.13 Principle of radar ranging

can be calculated based on the time difference between the signals received by two different position sensors.

Example 4.4 Radar Ranging The principle of radar ranging is illustrated in Fig. 4.13. The time difference between the emitted wave and the reflected wave can be obtained by correlation analysis. The distance of the target can be calculated with the equation shown in Fig. 4.13.

Example 4.5 Localization of Leakage Point in Pipelines Its principle is shown in Fig. 4.14. In order to detect the leakage location of the oil pipeline buried deep underground, acoustic sensors are placed at detection point 1 and detection point 2, respectively. The acoustic wave caused by oil leakage at the point K propagates along the pipe wall to both sides. Cross-correlation analysis is applied to the sound signals measured by the two acoustic sensors to find out the time delay of the two signals. When the transmission speed of the acoustic wave in the pipeline is known, the distance between the damaged position and centerline of two sensors can be determined:

$$S = \frac{1}{2} v \tau_0 \quad (4.22)$$

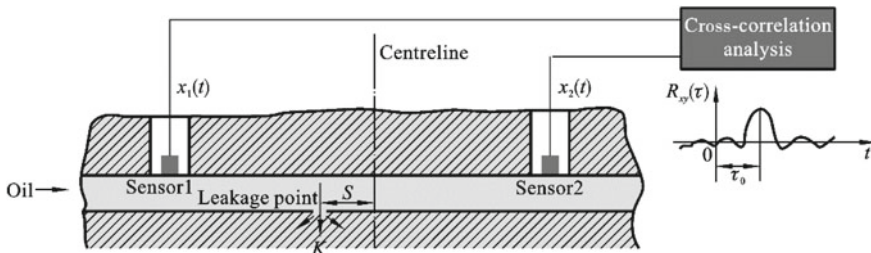


Fig. 4.14 Leakage point detection of pipelines

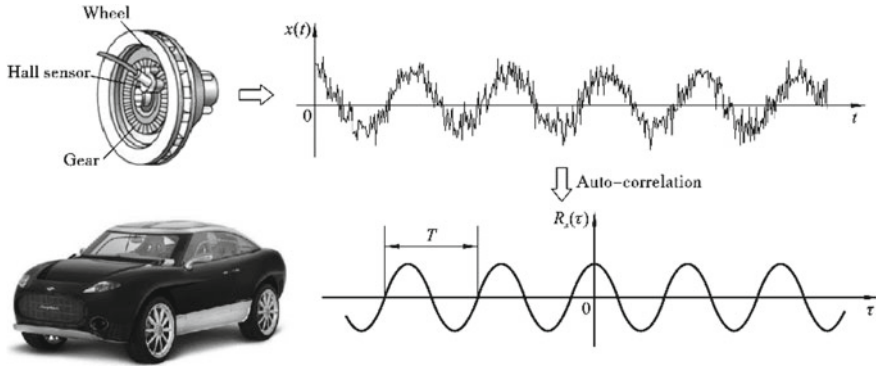


Fig. 4.15 Separation of noise signal in vehicle speed measurement

3. Separation of noise signal and periodic signal

In the measurement of rotation speed and natural vibration frequency of machines, the measured signal is theoretically a periodic signal. However, electromagnetic interference, environmental noise, thermal noise of the measurement instrument, may be superimposed into the measurement signal. From the properties of autocorrelation, we know that the autocorrelation of noise signal decays rapidly, while the autocorrelation function of the periodic signal is still a periodic signal of the same frequency. These properties can be used to separate periodic signals from the noises.

Example 4.6 Separation of Noise Signal in Vehicle Speed Measurement Figure 4.15 shows the vehicle speed measurement signal obtained by a Hall sensor. Due to the interference of noise, it is difficult to identify the period of the signal and the vehicle speed. As shown in Fig. 4.15, the autocorrelation $R_x(\tau)$ of the signal $x(t)$ is calculated, and the noises are suppressed. From the signal of $R_x(\tau)$, we can easily find the signal period and calculate the vehicle speed.

Example 4.7 Denoising by Autocorrelation in MATLAB The following MATLAB code is to reduce the noise in a signal by performing autocorrelation (Fig. 4.16).

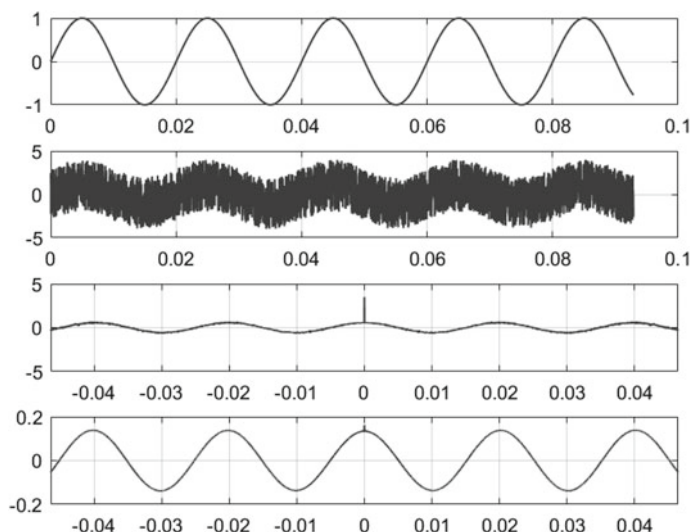


Fig. 4.16 Separating noise from the periodic signal in MATLAB

```

Fs=44100; N=4096;
dt=1.0/Fs; T=dt*N;
t=linspace(0,T,N);
x=sin(2*3.14*50*t);
subplot(4,1,1); plot(t,x,'linewidth',1);
grid on;
×1=x+3*rands(1,N);
subplot(4,1,2); plot(t,×1,'linewidth',1);
grid on;
r=xcorr(×1,'unbiased');
N1=length(r);
t1=linspace(-T,T,N1);
subplot(4,1,3); plot(t1,r,'linewidth',1);
grid on; xlim([-T/2,T/2]);
r1=xcorr(r,'unbiased');
N2=length(r1);
t2=linspace(-2*T,2*T,N2);
subplot(4,1,4); plot(t2,r1,'linewidth',1);
grid on; xlim([-T/2,T/2]);

```

Exercise

- 4.1 What is the definition of correlation function? What are the properties of correlation function?
- 4.2 What is the definition of convolution? What are the similarities and differences with correlation function?

- 4.3 Assume a signal $x(t)$ is the superposition of two cosine signals with different frequencies, $x(t) = A_1 \cos(\omega_1 t + \phi_1) + A_2 \cos(\omega_2 t + \phi_2)$. Find its autocorrelation function and plot the curve.
- 4.4 Find the cross-correlation function of a square wave and a sine wave of the same period.
- 4.5 What are the engineering applications of correlation analysis?

Chapter 5

Time–Frequency Domain Analysis



5.1 Motivations of Time–Frequency Domain Analysis

5.1.1 *Non-stationary Signals*

A signal is referred as non-stationary signal if its characteristics such as its mean value, variance, or center frequency, changes over time. Non-stationary signal is also known as time-varying signal.

Example 5.1 Chirp Signal The frequency of the chirp signal (Fig. 5.1) increases with time. Thus it is a non-stationary signal.

Example 5.2 Bell Sound Signal The variance of the bell sound signal (Fig. 5.2) becomes smaller. Thus it is a non-stationary signal.

5.1.2 *Drawbacks of Global Analysis of Non-stationary Signals*

Based on the characteristics of non-stationary signals, there are special requirements for its analysis. There are different drawbacks of applying the previous learned analyzing methods. For example, waveform analysis is only suitable for simple signals. For complex signals with multiple frequency components, frequency information cannot be displayed. As for the frequency domain analysis based on Fourier transform, it is only suitable for stationary signal analysis. When applied to non-stationary signals, only the global frequency information can be given, and the frequency information at a specific moment cannot be known. A comparison of FFT spectra of stationary signal and non-stationary signal is given in Fig. 5.3. The signal on top is the superposition of four sinusoids with frequencies of 5, 10, 25 and 50 Hz. The following two signals are non-stationary signals of single frequency

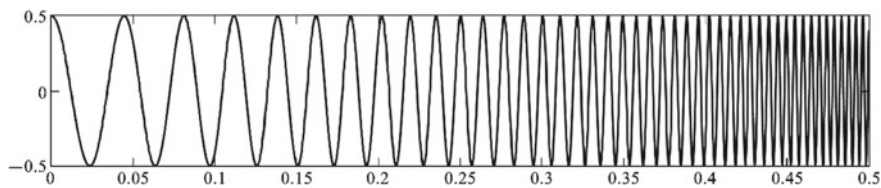


Fig. 5.1 Chirp signal waveform

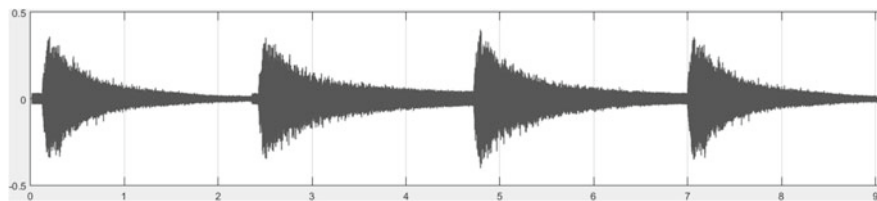


Fig. 5.2 Bell sound signal waveform

sinusoid appearing in different time span. As we can see, these three completely different signals have very similar spectrum. Although the spectra can show the global frequency components in the signals, the information in time domain (i.e. when the frequency component appear) is lost. For non-stationary signals, it is not enough to know which frequency components are included, we also need to know the time when each component appears, how the signal frequency changes with time, the instantaneous frequency and its amplitude at each moment. Therefore, the time–frequency domain analysis that can show the instantaneous frequency components is in need.

5.1.3 Time–Frequency Domain Analysis

In the analysis of stationary signals, time and frequency are two very important variables. Fourier transform is a bridge that links frequency domain and time domain. However, Fourier transform analyzes the signal globally, i.e. across the whole time span. Thus, it cannot show the frequency component at certain time and how frequency changes with time. In order to obtain these information, time–frequency domain analysis is needed, which maps the 1-D time domain signal into a 2-D image with time and frequency as the axes. The value of each pixel in the image represents the intensity of one frequency component appearing at a certain time. An example is given in Fig. 5.4. The time domain waveform of a chirp signal with frequency changing from 20 to 300 Hz within 1 s is shown in Fig. 5.4a. Its spectrum obtained by Fourier transform is shown in Fig. 5.4b, in which we can see the frequencies but not the change of frequency with time. The time–frequency domain representation

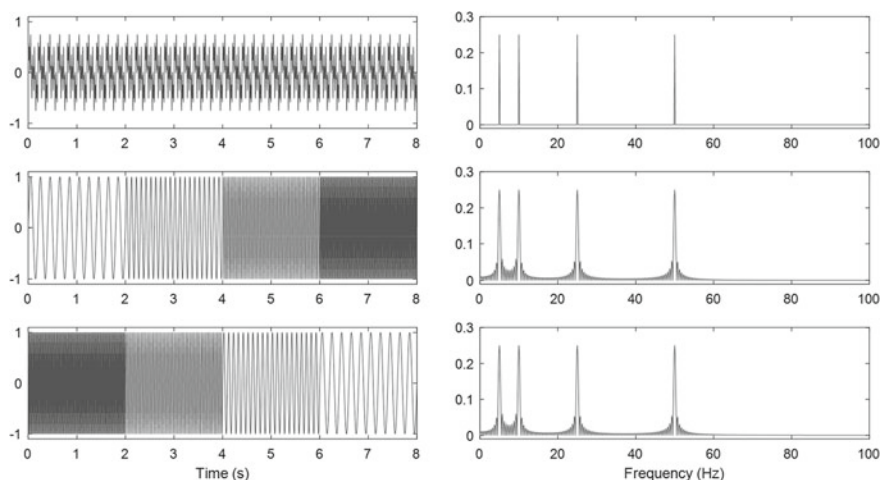


Fig. 5.3 Comparison of FFT spectra of stationary signal and non-stationary signals

is given in Fig. 5.4c, in which we can clearly see how the frequency component changes with time.

With time–frequency domain analysis, the distribution of energy in time and frequency can be simultaneously obtained. The commonly used time–frequency representations include linear and nonlinear types. The linear representations include short-time Fourier transform, continuous wavelet transform, etc., and the non-linear ones include Wigner-Ville distribution, Cohen distribution, etc.

5.2 Short-Time Fourier Transform

5.2.1 Basic Principle of Short-Time Fourier Transform

1. Basic principle

In order to overcome the lack of time resolution in Fourier transform, two methods are usually used to modify it to make it suitable for analyzing non-stationary signals. The first method is to observe a fragment of the signal through a window, i.e. multiplying the signal to be analyzed by a window function to truncate it. The duration of the fragment is short enough to be approximated as a stationary signal. In this way, the frequency components of the fragment can be locally analyzed by Fourier transform. Then the window is shifted to analyze other fragments of the signal, as shown in Fig. 5.5. The second method is to modify the base function used in the Fourier transform, i.e. sinusoidal functions, to a base function that is more concentrated in time and more dispersed in frequency.

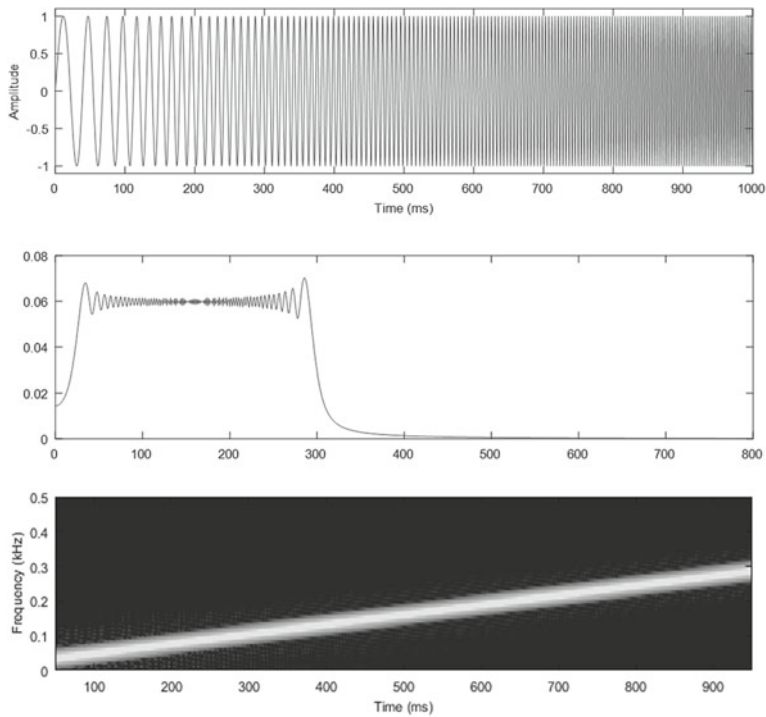


Fig. 5.4 Three methods to represent a chirp signal: **a** time domain; **b** frequency domain; **c** time–frequency domain

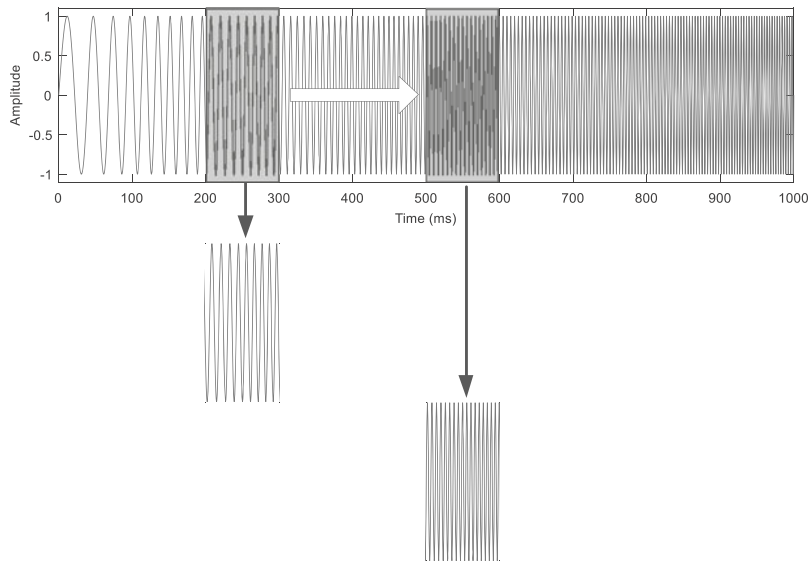


Fig. 5.5 Observe the signal through a window

2. Definition and property of short-time Fourier transform

The short-time Fourier transform (STFT) of signal $x(t)$ is defined as:

$$\text{STFT}\{x(t)\} \equiv X(\tau, \omega) = \int_{-\infty}^{\infty} x(t)w(t - \tau)e^{-j\omega t} dt \quad (5.1)$$

Equation (5.1) is actually the Fourier transform of $x(t) \cdot w(t - \tau)$. Thus, the operation of STFT corresponds to the following steps: (1) truncate the signal $x(t)$ by multiplying with a window function to get a segment of the signal centered at τ ; (2) implement Fourier transform to get the spectrum of the segment; (3) shift the window and repeat the frequency domain analysis.

The selection of window function will affect the truncation and hence affect the result of STFT. The commonly used windows include rectangular window, triangular window, Gaussian window, Hann window and Hamming window. Since it is better to focus the analysis at the time τ , a window function with rapid decay, such as Gaussian window and Hann window are preferred. But in some other cases, we need to consider the calculation efficiency. In this case, rectangular window is commonly used.

Fourier transform has the property of time delay and frequency shift. The time delay and frequency shift properties of the short-time Fourier transform are affected by the transform of base signal. According to Eq. (5.1), the following can be easily derived:

$$\tilde{x}(t) = x(t)e^{j2\pi f_0 t} \Rightarrow \tilde{X}(\tau, f) = X(\tau, f - f_0) \quad (5.2)$$

and

$$\tilde{x}(t) = x(t - t_0) \Rightarrow \tilde{X}(\tau, f) = X(\tau - t_0, f)e^{-j2\pi t_0 f} \quad (5.3)$$

Equations (5.2) and (5.3) shows that STFT retained the frequency shift and time delay property properties of $x(t)$.

Example 5.3 STFT in MATLAB The API function “spectrogram” in MATLAB can be called to perform STFT. A modulated chirp signal is analyzed by the following code. We can clearly see the change of frequency with time in the STFT result (Fig. 5.6).

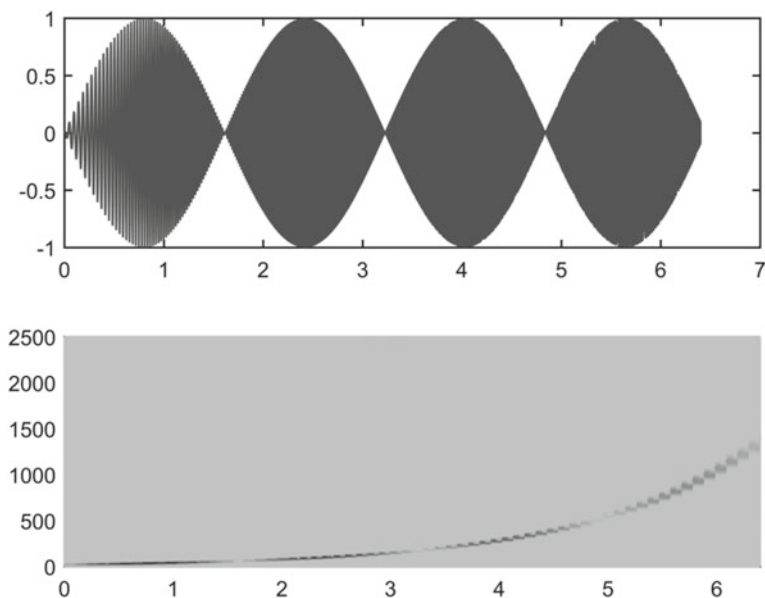


Fig. 5.6 STFT by MATLAB

```

Fs = 5120; N=32768;
dt=1.0/Fs; T=dt*N;
t=linspace(0,T,N);
x1=sin(2*pi*0.31*t);
x2=chirp(t,20,T,1500,'lo');
x3=x1.*x2;
subplot(2,1,1);
plot(t,x);
Z=spectrogram(x,1024,512);
P=sqrt(Z.* conj(Z));
[NN MM]=size(P);
X=linspace(0,Fs/2,NN);
Y=linspace(0,dt*N,MM);
subplot(2,1,2);
surf(Y,X,P); shading flat
view([0,0,1]);
xlim([0,6.4])
ylim([0,2560])

```

5.2.2 Time–Frequency Resolution of Short-Time Fourier Transform

The short-time Fourier transform maps a one-dimensional signal $x(t)$ into a two-dimensional function $X(\tau, f)$ in the time–frequency plane. The resolution of time and frequency will directly affect the accuracy of analysis. According to the definition of short-time Fourier transform, to have good time resolution, the window $w(t)$ should be as narrow as possible. However, to have good frequency resolution, the bandwidth of the window should be as narrow as possible, i.e. the corresponding window in time domain should be as wide as possible. Therefore, there is a contradiction between time resolution and frequency resolution. According to the Hersenberg’s uncertainty principle, if Δt and Δf are used to represent time resolution and frequency resolution, they satisfy the following inequality

$$\Delta t \cdot \Delta f \geq \frac{1}{4\pi} \quad (5.4)$$

The uncertainty principle prevents the result to have good resolution in both time domain and frequency domain. In practice, to get higher frequency resolution, time resolution needs to be sacrificed, or vice versa.

Two extreme cases are as follows.

1. Ideal time resolution

If we choose to have ideal time resolution, the window function should be infinitely narrow. Thus $\delta(t)$ function can be selected as the window, then

$$w(t) = \delta(t) \Rightarrow X(\tau, f) = x(\tau)e^{-j2\pi f\tau} \quad (5.5)$$

In this case, STFT degenerates to time domain function which loses frequency resolution.

2. Ideal frequency resolution

If we choose to have ideal frequency resolution, the window function should be constant $w(t) = 1$. Then, we get:

$$W(f) = \delta(f) \Rightarrow X(\tau, f) = X(f) \quad (5.6)$$

In this case, STFT degenerates to Fourier transform, which cannot provide any time resolution.

It can be seen that the improvement of time resolution leads the reduction of frequency resolution, and the improvement frequency resolution will inevitably lead to the sacrifice of time resolution. Therefore, it is necessary to pay attention to these two aspects in the process of short-time Fourier transform.

5.2.3 Time–Frequency Domain Decomposition and Synthesis

Another way to describe STFT is: represent $x(t)$ as the linear superposition of shifted base signal $g(t)$:

$$x(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} T_x(\tau, f) [g(t - \tau) e^{j2\pi f t}] d\tau df \quad (5.7)$$

According to this equation, we can also consider $x(t)$ is decomposed with $g_{\tau, f}(t) = g(t - \tau) e^{j2\pi f t}$. $T_x(t, f)$ is the coefficient of the decomposition, which describes the intensity of $x(t)$ at a point (t, f) in time–frequency domain. $T_x(t, f)$ can be also treated as the STFT result:

$$T_x(\tau, f) = X(\tau, f) = \int_{-\infty}^{\infty} x(t) w(t - \tau) e^{-j2\pi f t} dt \quad (5.8)$$

For Eq. (5.8) to hold, the selected window function should satisfy: $\int g(t) w(t) dt = 1$.

The time–frequency expansion in Eq. (5.7) exists for any signal with finite energy. Substitute Eq. (5.8) into Eq. (5.7), we get:

$$x(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} X(\tau, f) g(t - \tau) e^{j2\pi f t} d\tau df \quad (5.9)$$

This equation shows the recovery from the STFT result to time domain signal $x(t)$.

Equation (5.8) is called the analysis equation of STFT, and Eq. (5.9) is called the synthesis equation of STFT. The synthesis equation can be regarded as the inverse transform of the analysis equation.

5.2.4 Discrete Short-Time Fourier Transform

Sample the continuous STFT at intervals of (nT, kF) , where $T > 0$ and $F > 0$ are the sampling periods for time and frequency respectively, n and k are integers. The analysis equation Eq. (5.8) can be discretized as:

$$X[m, k] = \sum_{n=0}^{N-1} x[n + m] w[n] e^{-j \cdot 2\pi \cdot \frac{k}{N} n} \quad (5.10)$$

Take the squares of STFT coefficients, we can get the power spectrum, also known as spectrogram, of STFT. It reflects the power spectral density in the time–frequency plane. For continuous form, the power spectrum is:

$$\text{spectrogram}\{x(t)\} = |X(\tau, f)|^2 \quad (5.11)$$

And in discrete form is:

$$\text{spectrogram}\{x[n]\} = |X[m, k]|^2 \quad (5.12)$$

The discrete short-time Fourier transform algorithm includes two steps of processing, namely signal segmentation and spectrum estimation. Among them, the segmentation of the signal is to obtain the discrete short sequence through the sliding analysis window; the spectrum estimation is to estimate the spectrum of each short sequence.

Example 5.4 STFT of a Chirp Signal in MATLAB The time domain waveform and frequency spectrum of a chirp signal is shown in Fig. 5.7a, b. Its spectrogram can be obtained in MATLAB and the result is shown in Fig. 5.7c. The MATLAB code is listed below.

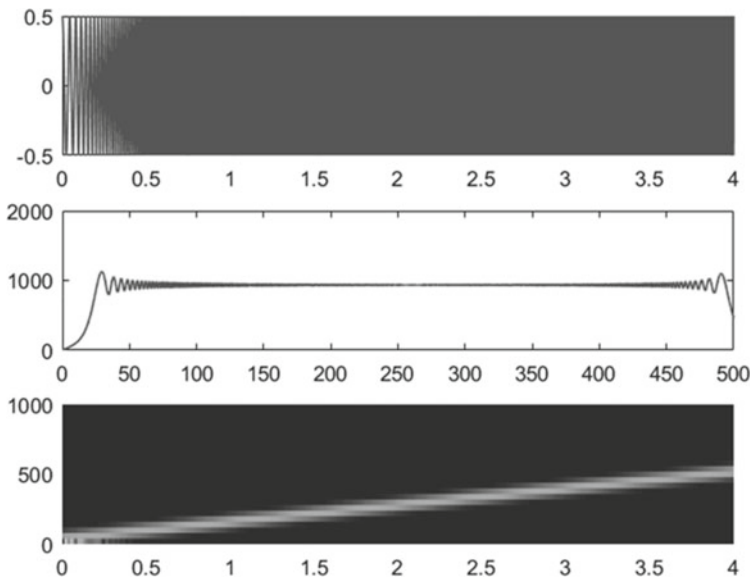


Fig. 5.7 Time domain waveform, power spectrum and STFT of a chirp signal

```

Fs=40960; dt=1.0/Fs; T=4; N=T/dt;
x=linspace(0,T,N);
y =0.5*chirp(x,20,T,500,'li');
t=linspace(0,T,N); subplot(3,1,1);
plot(t,y,'LineWidth',1);
z=fft(y); A1=abs(z);
M=length(A1);
f=linspace(0,Fs/2,M/2); subplot(3,1,2);
plot(f,A1(1:M/2),'linewidth',1);
xlim([0,500]);
Z=spectrogram(y,1024,512);
P=sqrt(Z.* conj(Z));
[NN MM]=size(P);
X=linspace(0,Fs/2,NN);
Y=linspace(0,dt*N,MM);
subplot(3,1,3);
surf(Y,X,P); shading flat
view([0,0,1]);
ylim([0,1000]);
zlim([0,200]);

```

5.2.5 *Methods to Improve Estimation of Spectrogram*

As mentioned earlier, in the short-time Fourier transform, the time–frequency resolution of signal is determined by the width of the time window and frequency window, and there is a contradiction between the two. To increase the frequency resolution, the time window width must be increased, which will cause the reduction of time resolution; conversely, to increase the time resolution, the time window width must be reduced, which makes each short sequence containing less amount of data. When the amount of data is small, the FFT analysis will inevitably have problems such as low frequency resolution and spectral leakage. In order to alleviate this contradiction, the maximum entropy spectrum estimation algorithm in modern spectrum analysis methods can be used to replace FFT for spectrum estimation.

It has been proved that the maximum entropy spectrum is equivalent to the autoregressive (AR) spectrum of the autoregressive model. The short-time AR spectrum is obtained through segmented time series modeling, and its estimation accuracy is higher than that of Fourier spectrum, and it has the advantages of smooth spectrum and sharp peaks. It does not have problems such as windowing and is especially suitable for short sequence analysis. At present, the parameter estimation algorithm of the AR model is very mature. After obtaining the AR model parameters of the short sequence at time t , the local AR spectrum at that point can be directly calculated to obtain the short-time AR spectrum.

5.3 Wavelet Transform

Wavelet transform was first proposed by French geophysicist Morlet in the early 1980s as a mathematical tool for analysis of geophysical signals. After years of development, wavelet transform has not only made breakthroughs in theory and methods, but also in applications of signal and image analysis, geophysical signal processing, computer vision and coding, speech synthesis and analysis, signal abnormality detection and spectrum estimation. It has even been widely used in fractal and chaos theory. In principle, the place where Fourier transform is traditionally used can now be replaced by wavelet transform. The advantage of wavelet transform over Fourier transform is that it has good localization properties in both time domain and frequency domain. In addition, due to the use of gradually refined time-domain or spatial-domain sampling resolution for high-frequency components, it can focus on any detail of the signal. In this sense, it is seen as a digital microscope, and it is a milestone in the history of Fourier analysis.

5.3.1 Wavelet Transform and Short-Time Fourier Transform

Short-time Fourier transform and wavelet transform are the two most important transforms in linear time–frequency analysis. The short-time Fourier transform maps a one-dimensional signal $x(t)$ into a two-dimensional function in the time–frequency plane (t, f) . Its main advantage is: if the energy of the signal is concentrated in a given time interval $[-T, T]$ and frequency interval $[-\Omega, \Omega]$, the energy can be converted into a localized region of $[-T, T] \times [-\Omega, \Omega]$ after the transform. For time and frequency intervals where there is not much energy in the signal, the result of the transform is close to zero, and this is very advantageous for extracting information in the signal. The main drawback is: because the same window is used for all frequencies, the resolution of the analysis is the same in all local areas of the time–frequency plane, as shown in Fig. 5.8. If there are components with short duration and high frequency in the signal, then the short-time Fourier transform is not very effective. Although narrowing the time window and reducing the sampling period can obtain more accurate information, but subject to the uncertainty principle, it is impossible to have high resolution in both time and frequency.

The time resolution of wavelet transform is adjustable for different frequencies. It gives narrow windows for high frequency components and wide windows for low frequency components, as shown in Fig. 5.9. Wavelet transform views the signal with different scales (or resolutions) in signal analysis. This multi-resolution or multi-scale view is the basic characteristic of wavelet transform. Its purpose is to see both the whole signal and details of the signal, which is like to see the whole forest and details of trees simultaneously. Therefore, wavelet transform is known as a mathematical microscope. This is why the wavelet transform is superior to the short-time Fourier transform.

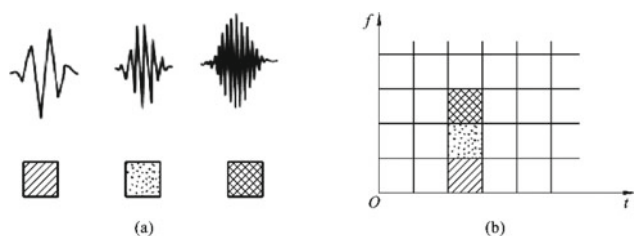


Fig. 5.8 Resolution of time and frequency of STFT

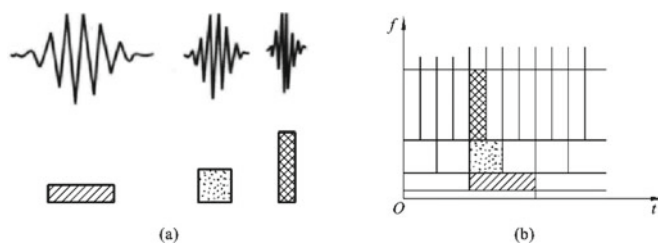


Fig. 5.9 Resolution of time and frequency of WT

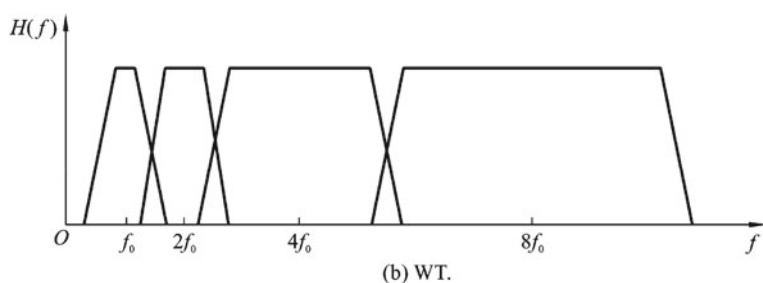
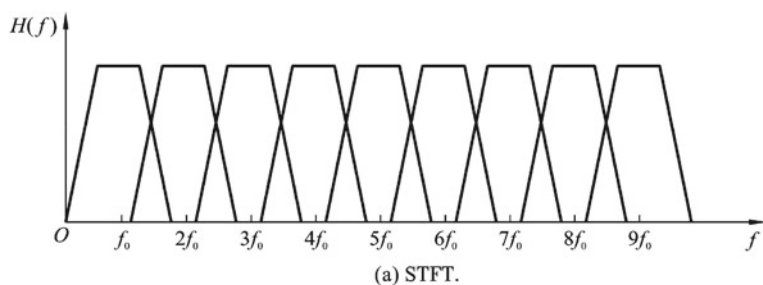


Fig. 5.10 Bandwidth of the two transforms: **a** STFT; **b** WT

If the wavelet transform and short-time Fourier transform are explained from the point of view of filtering. Intuitively, when the wavelet transform is regarded as filters, the time resolution must increase with the increase of the center frequency of the analysis filter. Therefore, the bandwidth Δf of the band-pass filter is proportional to the center frequency f , namely:

$$\frac{\Delta f}{f} = C \quad (5.13)$$

where C is a constant. As for short-time Fourier transform, the bandwidth of the band-pass filter has nothing to do with the analysis frequency or the center frequency f of the filter. As shown in Fig. 5.10, the bandwidth of the band-pass filter of short-time Fourier transform is uniformly and regularly distributed on the frequency axis (i.e. constant bandwidth), while the that of the wavelet transform is expanded on the frequency axis with a logarithmic scale (i.e. constant relative bandwidth).

5.3.2 Continuous Wavelet Transform

Let symbols \mathbb{Z} and \mathbb{R} represent the set of integers and real numbers, $L^2(\mathbb{R})$ and $L^2(\mathbb{R}^2)$ represent square-integrable one-dimensional and two-dimensional functions, i.e. the vector space of $f(x)$ and $f(x, y)$. $l^2(\mathbb{Z})$ is the vector space of the summable sequence of squares, namely

$$l^2(\mathbb{Z}) = \left\{ (a_i)_{i \in \mathbb{Z}} : \sum_{i=-\infty}^{\infty} |a_i|^2 < \infty \right\} \quad (5.14)$$

If $\psi \in L^2(\mathbb{Z})$ satisfy:

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < +\infty \quad (5.15)$$

where $\hat{\psi}(\omega)$ is the Fourier transform of $\psi(t)$, then $\psi(t)$ is called a mother wavelet.

Based on the mother wavelet, a series of wavelets can be obtained by scaling and shifting:

$$\psi_{b,a}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \quad (a, b \in \mathbb{R}, a \neq 0) \quad (5.16)$$

where $\psi_{b,a}(t)$ is sometimes called child wavelet. Large a indicates a widened wavelet, and corresponds to a low frequency function; small a indicates a shrunk wavelet and corresponds to a high frequency function.

For a non-stationary signal $x(t) \in L^2(\mathbb{R})$, its continuous wavelet transform is defined as:

$$(W_\psi x)(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^*\left(\frac{t-b}{a}\right) dt \quad (5.17)$$

where $a, b \in L^2(\mathbb{R})$, $a \neq 0$, $\psi^*\left(\frac{t-b}{a}\right)$ is the conjugation of $\psi\left(\frac{t-b}{a}\right)$. Equation (5.17) can be also regarded as inner product of $x(t)$ and function $\psi_{b,a}(t)$:

$$(W_\psi x)(b, a) = \langle x(t), \psi_{b,a}(t) \rangle \quad (5.18)$$

The condition for existing mother wavelet $\psi(t)$ is defined in Eq. (5.15), from which, we can infer that $\hat{\psi}$ is a continuous function. Thus, $\hat{\psi}(0) = 0$ can be inferred from the admissibility criterion in Eq. (5.15), or equivalently expressed as:

$$\int_{-\infty}^{\infty} \psi(t) dt = 0 \quad (5.19)$$

And because $\psi(t)$ it is a square integrable function, the wavelet function must be a short wavelet that oscillates and decays quickly, and cannot be a periodic function. This is why it is called a “wavelet”.

The continuous wavelet transform $(W_\psi x)(b, a)$ defined by Eq. (5.17) depends on two parameters a and b . a is called the dilation factor that makes the mother wavelet function stretch or compress on the abscissa axis, and b is called the translation factor that makes the function shift on the abscissa axis. Assuming that the time window center and the time window width of the mother wavelet are t^* and Δ_ψ respectively, then the function $\psi_{b,a}$ is a window function with the center of $b + at^*$ and the radius of $a\Delta_\psi$. Therefore, the continuous wavelet transform shown in Eq. (5.16) gives the local information of a signal in the time window $[b + at^* - a\Delta_\psi, b + at^* + a\Delta_\psi]$. This window width changes with a .

If we consider

$$\begin{aligned} \frac{1}{2\pi} \hat{\psi}_{b,a}(\omega) &= \frac{1}{2\pi\sqrt{a}} \int_{-\infty}^{\infty} e^{-j\omega t} \psi\left(\frac{t-b}{a}\right) dt \\ &= \frac{\sqrt{a}}{2\pi} e^{-j\omega b} \hat{\psi}(a\omega) \end{aligned} \quad (5.20)$$

and suppose that the center and radius of the window function $\hat{\psi}$ are ω^* and $\Delta_{\hat{\psi}}$, and let

$$\eta(\omega) = \hat{\psi}(\omega + \omega^*) \quad (5.21)$$

then a window function with a center at the origin and a radius equal to $\Delta_{\hat{\psi}}$ is obtained. According to Eqs. (5.18) and (5.20) and Parseval's theorem, we get:

$$(W_{\psi}x)(b, a) = \frac{\sqrt{a}}{2\pi} \int_{-\infty}^{\infty} \hat{x}(\omega) e^{jb\omega} \eta * \left[a \left(\omega - \frac{\omega^*}{a} \right) \right] d\omega \quad (5.22)$$

Because the window function $\eta \left[a \left(\omega - \frac{\omega^*}{a} \right) \right] = \eta(a\omega - \omega^*) = \hat{\psi}(a\omega)$ has radius $\frac{1}{a} \Delta_{\hat{\psi}}$, thus Eq. (5.22) indicates that except for having a multiplication factor $\sqrt{a}/2\pi$ and a linear phase shift $e^{jb\omega}$, the continuous wavelet transform can also give the local information of the signal in the frequency window $\left[\frac{\omega^*}{a} - \frac{1}{a} \Delta_{\hat{\psi}}, \frac{\omega^*}{a} + \frac{1}{a} \Delta_{\hat{\psi}} \right]$.

If we regard ω^*/a as frequency variable ω , then we can regard t - ω plane as time-frequency plane. The rectangle window in time-frequency plane is:

$$\left[b + at^* - a\Delta_{\psi}, b + at^* + a\Delta_{\psi} \right] \times \left[\frac{\omega^*}{a} - \frac{1}{a} \Delta_{\hat{\psi}}, \frac{\omega^*}{a} + \frac{1}{a} \Delta_{\hat{\psi}} \right]$$

Its width is $2a\Delta_{\psi}$. In this time-frequency window, when $|a|$ becomes larger, the time window becomes wider and the frequency window becomes narrower; when $|a|$ becomes smaller, the time window becomes narrower and the frequency window becomes wider. The product of the time window and the frequency window is a certain value, regardless of a and b , and satisfies the uncertainty principle. This indicates that the wavelet transform has a good resolution in the frequency domain for low-frequency signals, and has good resolution in the time domain for high-frequency signals. Therefore, wavelet transform is a time-frequency analysis method with the ability of zooming.

For the continuous wavelet transform of any finite energy signal, the original signal can be reconstructed. The inverse transform of continuous wavelet transform can be defined by the following formula:

$$x(t) = \frac{1}{C_{\psi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [(W_{\psi}x)(b, a)] \psi_{b,a}(t) \frac{da}{a^2} db \quad (5.23)$$

In signal analysis, if only the positive frequency ω is considered, and the frequency variable ω is a positive constant multiple of the reciprocal of the parameter a , $\omega = \omega^*/a$, then only positive a is considered. Therefore, when reconstructing x from the wavelet transform of x , only the value $(W_{\psi}x)(b, a)$, $a > 0$ is used. There should be a restriction on mother wavelet:

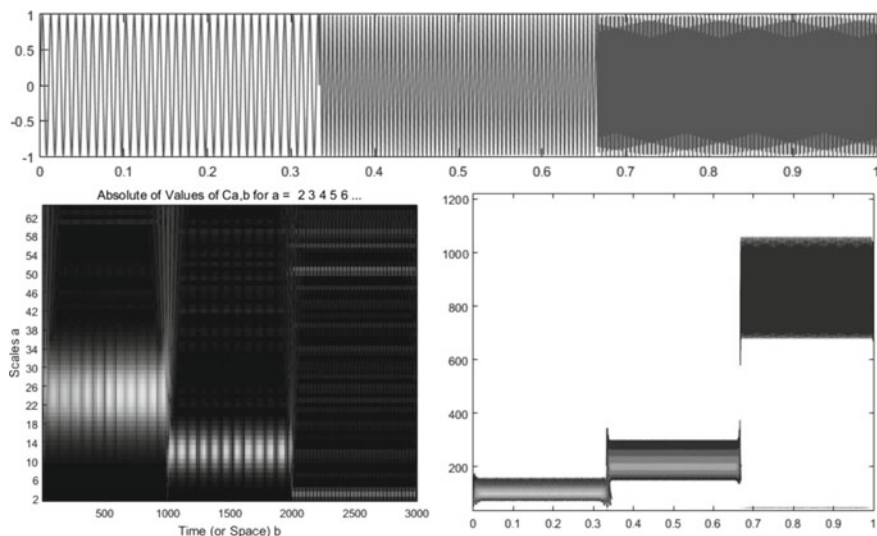


Fig. 5.11 CWT by MATLAB

$$\int_0^{\infty} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega = \int_0^{\infty} \frac{|\hat{\psi}(-\omega)|^2}{\omega} d\omega = \frac{1}{2} C_{\psi} < \infty \quad (5.24)$$

Example 5.5 Performing CWT in MATLAB The following MATLAB code is to analyze a piecewise function with changing frequency by continuous wavelet transform. The result is shown in Fig. 5.11.

```

Fs=3000; dt=1.0/Fs;
N=1000; T=dt*N;
t=linspace(0,T,N);
x1=sin(2*pi*100*t);
x2=sin(2*pi*200*t);
x3=sin(2*pi*800*t);
x=[x1 x2 x3];
N2=length(x);
T2=dt*N2;
t2=linspace(0,T2,N2);
plot(t2,x,'linewidth',1);
figure;
sc=2:64;
cfs=cwt(x,sc,'morl','plot');
figure;
pfreq=scal2frq(sc,'morl',1/Fs);
contour(t2,pfreq,abs(cfs));

```

5.3.3 Applications of Wavelet Analysis

Wavelet analysis is a rapidly developing new field in mathematics. The development is both in theory and application. In fact, the application fields of wavelet analysis are very wide, including: many subjects in the field of mathematics, signal analysis, image processing, quantum mechanics, theoretical physics, military electronics, classification and recognition, artificial synthesis of music and language, medical imaging and diagnosis, seismic exploration data processing and fault diagnosis of large machinery. In mathematics, it has been used in numerical analysis, fast numerical methods, curve and surface construction, differential equation solving, and control theory. In signal analysis, it is used for filtering, denoising, compression and transmission. In image processing, it is used for image compression, classification, identification and decontamination. In medical imaging, it is used to reduce the imaging time and improve the image quality in ultrasonic testing, CT and MRI.

Wavelet analysis for signal and image compression is an important aspect of wavelet analysis applications. It is characterized by high compression ratio, fast compression speed and can keep the characteristics of the signal and image after compression. Wavelet analysis can also resist interference in transmission. There are many compression methods based on wavelet analysis, and effective ones are the wavelet packet optimal basis method, the wavelet domain texture model method, the wavelet transform zero tree compression, and the wavelet transform vector compression.

Wavelet analysis is also widely used in signal analysis. It can be used for boundary processing and filtering, time–frequency analysis, signal-to-noise separation and extraction of weak signals, fractal index solving, signal identification and diagnosis, and multi-scale edge detection.

Application in engineering technology and other aspects includes computer vision, computer graphics, curve design, turbulence, long-distance cosmic research and biomedicine.

The prospects of wavelet analysis are as follows:

- (1) Transient signals or sudden changes in images often contain important fault information. For example, mechanical faults, power system faults, abnormalities in EEG and ECG, the location and shape of underground targets, etc., all correspond to abnormalities in measurement signal. Although these problems occur in different backgrounds, they can all be attributed to the problem of extracting the position of the abnormality in the signal. For an image, the sharply changing point usually corresponds to the edge that represents the image structure. If the edge is obtained, the basic features of the image if also obtained. Therefore, the application of wavelet analysis in fault detection and multi-scale edge extraction has a wide range of application.
- (2) The combination of neural network and wavelet analysis, and the combination of fractal geometry and wavelet analysis are one of the hotspots of the state-of-art research. Without the wavelet theory, the neural network based intelligent processing technology, fuzzy calculation, neural network based evolutionary

calculation cannot have good breakthrough. The research of nonlinear science is calling for wavelet analysis. In the future, nonlinear wavelet analysis may be an ideal tool for solving nonlinear scientific problems.

- (3) Wavelet analysis is used for data or image compression, and most of them are performed on static images. For network-oriented moving image compression, discrete cosine transform (DCT) and motion compensation (MC) have been used for a long time as coding technology. However, this method has two main drawbacks: blocking effect and mosquito noise. The multi-scale analysis of wavelet analysis can not only overcome the above-mentioned problems, but also can obtain the image outline on the coarse scale, and decide whether fine images should be transmitted to improve the image transmission speed. Therefore, the study of wavelet analysis algorithms for network-oriented low-rate image compression is worth exploring.
- (4) The two-dimensional and high-dimensional mother wavelets being used are mainly separable wavelets. The theories on constructions, properties and applications of inseparable two-dimensional and high-dimensional wavelets lack study. Perhaps the research of vector wavelet and high-dimensional wavelet can create a new world for the application of wavelet analysis.

Exercise

- 5.1 What are the differences between stationary and non-stationary signals?
- 5.2 What are the advantages of time–frequency domain analysis?
- 5.3 Design a MATLAB GUI to perform time–frequency domain analysis of audio signals.
- 5.4 Design a MATLAB GUI to perform time–frequency domain analysis of chirp signal and sinusoids with multi-frequencies.

Chapter 6

Digital Filters



6.1 Concept of Filtering

Signals inevitably contain some undesirable interference components, such as high frequency electromagnetic noise in the environment, thermal drifting noise, or acoustical noise you hear during a class. To improve the quality of signal, we have to try to remove or reduce the noise in signals. Filtering is a conventional way for denoising. Filtering is to remove or greatly suppress some undesired frequency components in the signal to get a relatively pure signal. As shown in Fig. 6.1, a distorted sinusoidal signal is difficult to be analyzed in time domain, but it is clear to see that there are three frequency components from the frequency domain. We can try to remove the high frequency noises to improve the signal quality.

Filter can be used in measurement instrument to suppress or eliminate noise and extract useful signal. Filter is a frequency selection device that can pass specific frequency components in the signal while greatly attenuating other frequency components. According to signal to be processed, it can be divided into analog filter and digital filter. According to the frequency band of the output signal, it is can be divided into low-pass, high-pass, band-pass, and band-stop filters. According to the electronic components used, it is divided into passive and active filters.

Example 6.1 Filtering the Following Signal with Different Filters

$$x(t) = 10 \sin(2\pi 50t) + \frac{10}{2} \sin(2\pi 100t) + \frac{10}{3} \sin(2\pi 150t) + \frac{10}{4} \sin(2\pi 200t)$$

The signal waveform and spectrum are shown in Fig. 6.2.

After the signal passed low-pass, high-pass, band-pass, and band-stop filters, the time domain waveform and frequency domain spectrum are shown in Fig. 6.3.

The characteristics of filters are usually represented by the frequency response $H(j\omega)$. Since $H(j\omega)$ is a complex number, it can be further divided as amplitude response $|H(j\omega)|$ and phase response $\angle H(j\omega)$. The amplitude response is also referred as magnitude frequency response, which shows the ratio of output signal

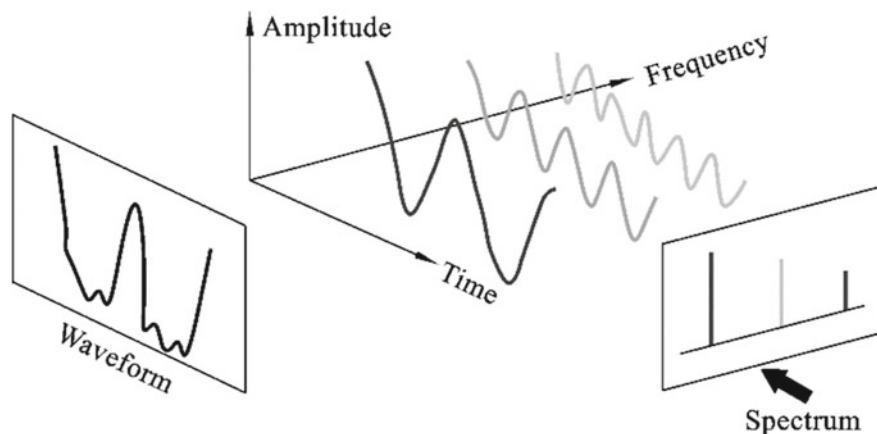


Fig. 6.1 Decomposition of complex signal

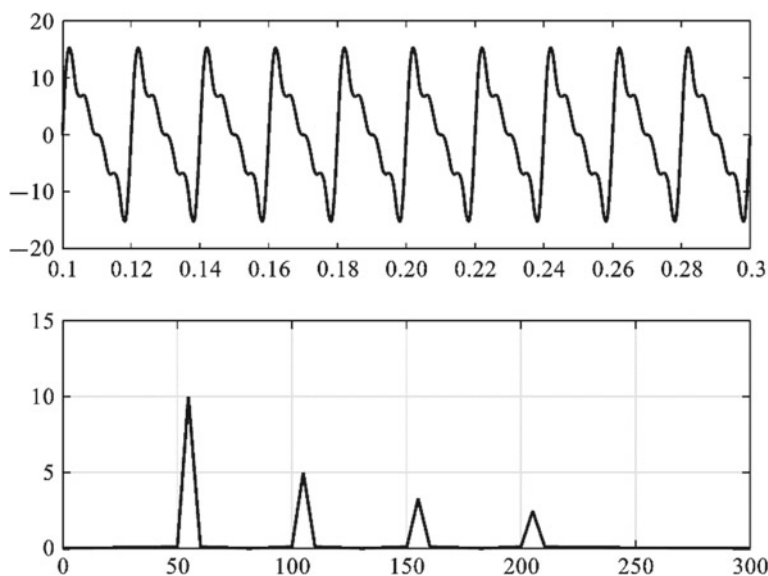


Fig. 6.2 Time domain waveform and frequency spectrum of a signal with multiple frequencies

amplitude to input signal amplitude at each frequency. The amplitude responses for the commonly used peak filter, notch filter, comb peak filter and comb notch filter are shown in Fig. 6.4.

Low-pass and high-pass filters are the two fundamental forms of filters. Other filters can be considered as combinations of these two types. For example, a band-pass filter is a serial connection of a low-pass filter and a high-pass filter. The cutoff frequency low-pass filter f_L and high-pass filter f_H should satisfy: $f_L > f_H$ (Fig. 6.5).

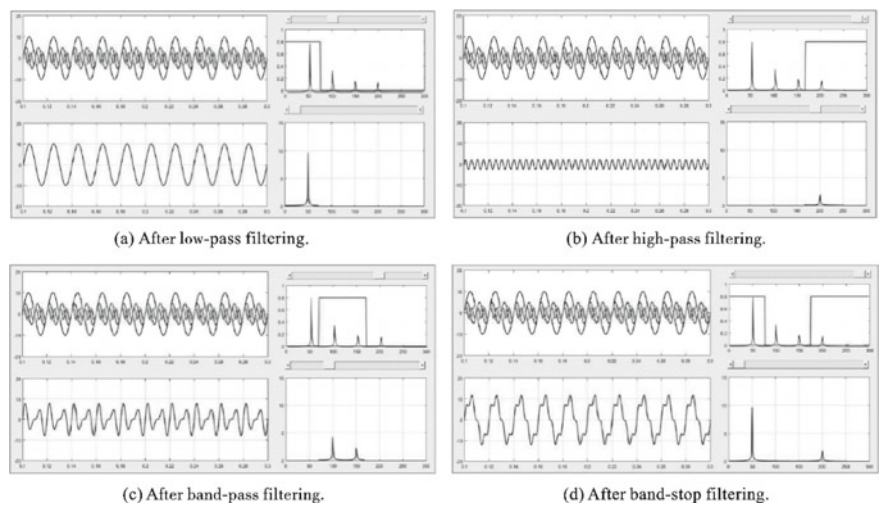


Fig. 6.3 Signals after filtering

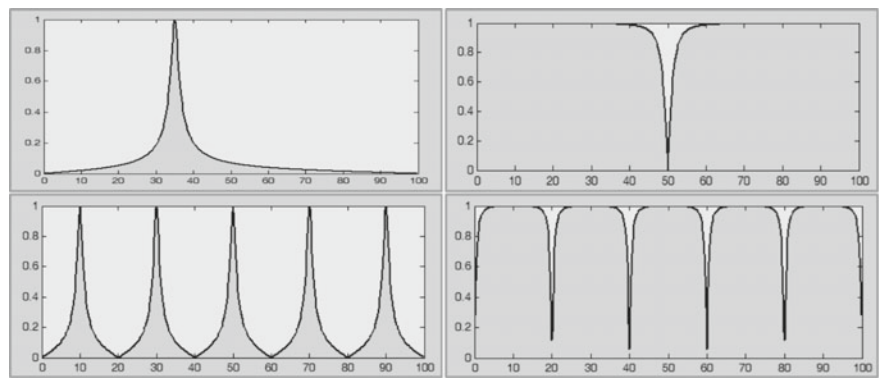


Fig. 6.4 Peak filter, notch filter, comb peak filter and comb notch filter

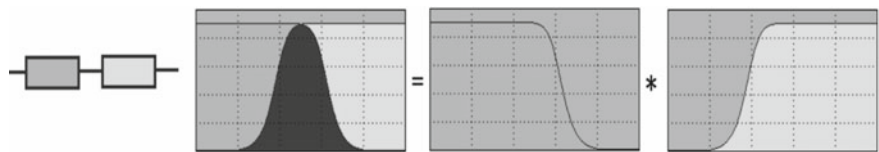


Fig. 6.5 Decomposition of band-pass filter

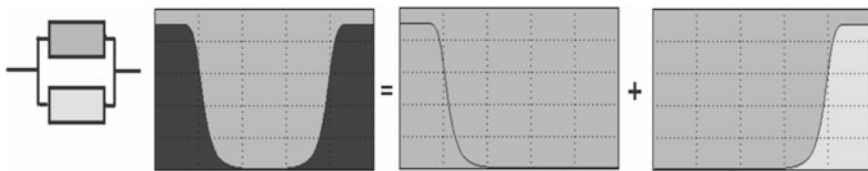


Fig. 6.6 Decomposition of band-stop filter

Band-stop filter is a low-pass filter and a high-pass filter connected in parallel. Their cutoff frequencies should satisfy: $f_L < f_H$ (Fig. 6.6).

6.1.1 Ideal Filters

1. Model

Ideal filter is an idealized model defined according to certain characteristics of the filter. It is a physically unrealizable filter, but it is helpful to understand the transfer characteristics of the filter. And some conclusions derived from it can be used as the basis for the analysis of the actual filter.

In the frequency domain, an ideal low-pass filter has rectangular amplitude response and linear phase response, as shown in Fig. 6.7. Its frequency response function, amplitude response and phase response are respectively:

$$H(f) = A_0 e^{-j2\pi f \tau_0} \quad (6.1)$$

$$|H(f)| = \begin{cases} A_0 & -f_c < f < f_c \\ 0 & \text{otherwise} \end{cases} \quad (6.2)$$

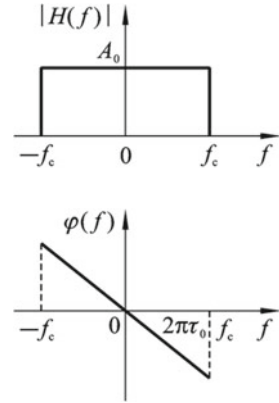
$$\varphi(f) = -2\pi f \tau_0 \quad (6.3)$$

The ideal low-pass filter passes the frequency components below the cut-off frequency f_c in the signal without any distortion and completely removes the frequency components above f_c .

2. Impulse Response

According to the transfer characteristics of the linear system, when the δ function passes through the ideal filter, its impulse response function $h(t)$ should be the inverse Fourier transform of the frequency response function $H(f)$, thus:

Fig. 6.7 The amplitude response and phase response of an ideal low-pass filter



$$\begin{aligned}
 h(t) &= \int_{-\infty}^{\infty} H(f) e^{j2\pi f t} df \\
 &= \int_{-f_c}^{f_c} A_0 e^{-j2\pi f \tau_0} e^{j2\pi f t} df \\
 &= 2A_0 f_c \frac{\sin 2\pi f_c (t - \tau_0)}{2\pi f_c (t - \tau_0)} \\
 &= 2A_0 f_c \text{sinc}[2\pi f_c (t - \tau_0)]
 \end{aligned} \tag{6.4}$$

The waveform of impulse response function $h(t)$ is shown in Fig. 6.8. This is a sinc function with its peak at τ_0 .

The following conclusions can be drawn:

- (1) when $t = \tau_0$, we get $h(t) = 2A_0 f_c$, where τ_0 is called the time delay, i.e. the time lag between the response signal and the excitation signal.
- (2) when $t = \tau_0 \pm n/2f_c$ ($n = 1, 2, \dots$), we get $h(t) = 0$, which means the impulse response function is a periodic function
- (3) when $t \leq 0$, we get $h(t) \neq 0$, which means the response to excitation signal $\delta(t)$ appears before the time it happens ($t = 0$).

The above analysis shows that the ideal filter is unrealizable. As we can see from the waveform of $h(t)$, the response to excitation signal $\delta(t)$ appears before the time

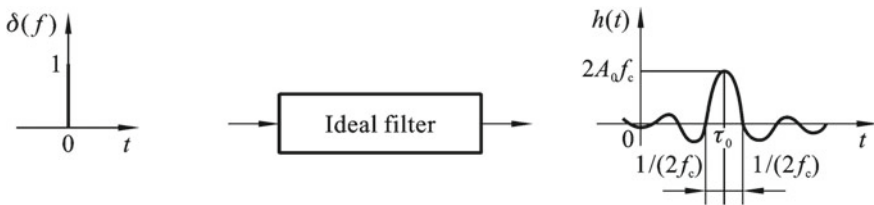


Fig. 6.8 The impulse response of an ideal filter

it happens. Of course this is unrealizable, since no filter could have the ability of “foreseen”. It can be inferred that the ideal high-pass, band-pass, and band-stop filters also do not exist. The amplitude response of an actual filter is unlikely to have a sharp change of right angle at the cutoff frequency. Thus, an actual filter cannot completely cut off at a certain frequency. Theoretically, the amplitude response of an actual filter will extend to infinity $|f| \rightarrow \infty$, so a filter can only greatly attenuate the frequency components outside the pass band, but cannot completely stop it.

3. Step Function Response

Discussing the step response of an ideal filter can further understand the transfer characteristics of filters and establish the relationship between the rise time of the filter response and the bandwidth of the filter. If a step function

$$u(t) = \begin{cases} 1 & t > 0 \\ \frac{1}{2} & t = 0 \\ 0 & t < 0 \end{cases} \quad (6.5)$$

is fed into the a filter, the output $y_u(t)$ is the convolution of input $u(t)$ and impulse response $h(t)$:

$$\begin{aligned} y_u(t) &= h(t) * u(t) \\ &= 2A_0 f_c \text{sinc}[2\pi f_c(t - \tau_0) * u(t)] \\ &= 2A_0 f_c \int_{-\infty}^{\infty} \text{sinc}[2\pi f_c(t - \tau_0)u(t - \tau)]d\tau \\ &= 2A_0 f_c \int_{-\infty}^t \text{sinc}[2\pi f_c(t - \tau_0)]d\tau \\ &= A_0 \left[\frac{1}{2} + \frac{1}{\pi} \text{si}(y) \right] \end{aligned} \quad (6.6)$$

where $\text{si}(y)$ is a notation for sine integral and is defined as:

$$\text{si}(y) = \int_0^y \frac{\sin x}{x} dx \quad (6.7)$$

and

$$\begin{aligned} y &= 2\pi f_c(t - \tau_0) \\ x &= 2\pi f_c(\tau - \tau_0) \end{aligned} \quad (6.8)$$

According to Eq. (6.6), the response of the ideal low-pass filter to the unit step input can be obtained as shown in Fig. 6.9.

The following conclusions can be drawn from Eq. (6.6):

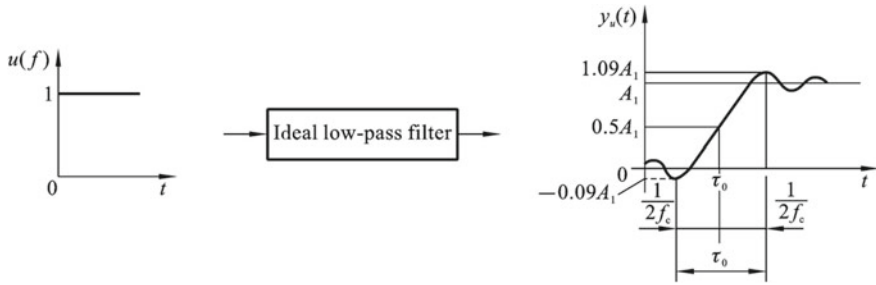


Fig. 6.9 The response of an ideal low-pass filter to a unit step input

- (1) when $t = \tau_0$, we get $y_u(t) = 0.5A_0$, where τ_0 is called the lag time of an ideal filter with unit step input;
- (2) when $t = \tau_0 + 1/2f_c$, we get $y_u(t) \approx 1.09A_0$, and when $t = \tau_0 - 1/2f_c$, we get $y_u(t) \approx -0.09A_0$. The time interval $(\tau_0 - 1/(2f_c), \tau_0 + 1/(2f_c))$ is called the time history of the filter response to the unit step. The interval length $\tau_d = 1/f_c$ is the rise time of the response to unit step. The bandwidth of the filter is $B = f_c$. Thus

$$\tau_d = \frac{1}{B} \quad (6.9)$$

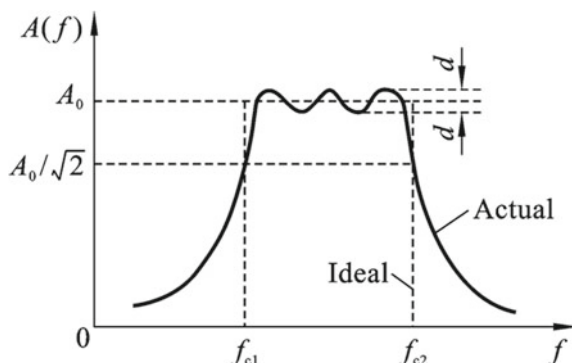
This equation shows that the rise time of the response of low-pass filter to unit step is inversely proportional to the bandwidth B , or in other words, the product of the rise time and the bandwidth is a constant. Its physical meaning can be explained as follows: The sudden change (discontinuity point) of the input signal contains rich high-frequency components. the low-pass filter attenuates the high-frequency components, and the result is that the signal waveform is smoothed. The wider the pass band, the less high-frequency components are stopped, which makes more energy to pass, so the rise time is shorter.

The bandwidth of a filter indicates its frequency resolution. The narrower the pass band, the higher the resolution. Therefore, the above conclusion is of great significance. It reminds us that the high resolution capability of a filter and the requirement of quick response are contradictory. If we want to use filters to select a very narrow frequency component from the signal (for example, if you want high-resolution spectrum analysis), we need enough time. If the time is not enough, errors and artifacts will inevitably occur.

6.1.2 Practical Filter

For an ideal filter, cut-off frequencies f_{c1} and f_{c2} are the only parameters to indicate its performance since the gain is constant between the cut-off frequencies, and is zero

Fig. 6.10 Amplitude responses of ideal band-pass and actual band-pass filters



outside, as shown by the dotted line in Fig. 6.10. As for the actual filter, as shown by the solid line, there is no sharp turning point, and the gain between cut-off frequencies is not a constant, thus we need more parameters to describe its performance. The main parameters are ripple amplitude, cutoff frequency, bandwidth, quality factor, octave selectivity, etc.

1. Ripple Amplitude D

Within a certain frequency range, the amplitude responses of the practical filter may exhibit ripple changes. Compared with the average value of the amplitude responses A_0 , the fluctuation range d should be as small as possible. Generally, it should be far less than -3 dB, that is $d \ll A_0/\sqrt{2}$.

2. Cut-Off Frequency f_c

The frequency corresponding to the value of $A_0/\sqrt{2}$ in the amplitude response curve is called the cut-off frequency of the filter. With the reference value of A_0 , $A_0/\sqrt{2}$ corresponds to a decay of -3 dB. At the cut-off frequency point, the signal power is attenuated by half.

3. Bandwidth B and Quality Factor Q

The frequency range between the upper and lower cut-off frequencies is called the filter bandwidth B , or -3 dB bandwidth. It has the unit of Hz. The bandwidth determines the filter's ability to separate adjacent frequency components in the signal, i.e. the frequency resolution. For band-pass filters, the ratio of the center frequency f_0 to the bandwidth B is usually called the quality factor Q of the filter. For example, a filter with center frequency of 500 Hz and bandwidth of 10 Hz, the Q factor is 50. The larger the quality factor, the higher the frequency resolution of the filter.

4. Octave Selectivity W

Outside the cut-off frequencies, the practical filter has a transition zone. The slope of the transition zone in the amplitude response curve indicates how fast the amplitude response decays. It determines the filter's ability to attenuate the frequency

components outside the bandwidth. Octave selectivity refers to the attenuation of the amplitude response between the upper cut-off frequency f_{c2} and $2f_{c2}$, or between the lower cut-off frequency f_{c1} and $2f_{c1}$, namely the attenuation value when the frequency changes by one octave. It is mathematically defined as:

$$W = -20 \lg \frac{A(2f_{c2})}{A(f_{c2})} \quad (6.10)$$

or

$$W = -20 \lg \frac{A\left(\frac{f_{c1}}{2}\right)}{A(f_{c1})} \quad (6.11)$$

Octave decay is presented with the unit dB/oct. Obviously, the faster the attenuation (i.e. the larger the W value), the better the filter selectivity. To express the attenuation rate far away from the cut-off frequency, we can also use the attenuation per 10 octave.

5. Filter Factor (or Rectangular Coefficient) λ

Another way to express filter selectivity is to use the ratio of -60 dB bandwidth to -3 dB bandwidth:

$$\lambda = \frac{B_{-60dB}}{B_{-3dB}} \quad (6.12)$$

Ideal filter has filter factor of $\lambda = 1$, while the commonly used actual filter has the filter factor between 1 to 5. For some filters, due to influence of electronic component (such as leakage resistance of capacitor, etc.), the attenuation factor of stopband can never reach -60 dB, and the selectivity is expressed by the ratio of the other bandwidth (such as -40 dB or -30 dB) to the -3 dB bandwidth.

6.1.3 Digital Filters

Digital filters are commonly used discrete systems. The input signal is a signal in a broad sense, it can be voltage, current, power, etc. In actual operation, we can also turn the input signal waveform into output, that is, invert the input and output, so as to achieve the modification the signal spectrum. The purpose of using a digital filter is to process the input signal waveform (or spectrum) to get desired output by attenuating some frequency components. In another word, it is to use digital method to transform the input sequence $x[n]$ to output $y[n]$ according to predetermined requirements to achieve the goal of changing the signal spectrum (Fig. 6.11).

Since digital filter is a discrete-time system, its time-domain input and output can be mathematically expressed with the equation:

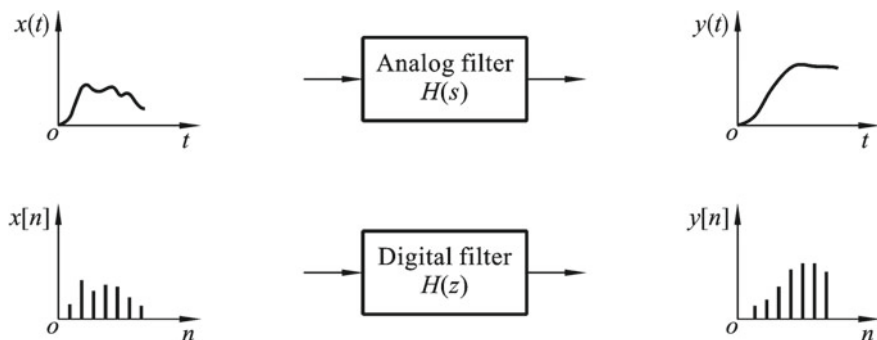


Fig. 6.11 Digital filter system

$$\begin{aligned}
 & a_0 y[n] + a_1 y[n-1] + \dots + a_{N-1} y[n-N+1] + a_N y[n-N] \\
 & = b_0 x[n] + b_1 x[n-1] + \dots + b_{M-1} x[n-M+1] + b_M x[n-M]
 \end{aligned} \quad (6.13)$$

Re-write Eq. (6.13) with the Σ sign, we get:

$$\sum_{k=0}^N a_k y[n-k] = \sum_{r=0}^M b_r x[n-r] \quad (6.14)$$

where $y[n]$ is the response and $x[n]$ is the input, a_0, a_1, \dots, a_N and b_0, b_1, \dots, b_M are constants, N and M are respectively the highest order of $y[n]$ and $x[n]$.

Suppose the impulse response sequence of the digital filter is $\{h[0], h[1], h[2], \dots\}$, then the input and output of the system can be written in discrete convolution form:

$$y[n] = h[n] * x[n] = \sum_{m=0}^{\infty} h[m] x[n-m] \quad (6.15)$$

The response in frequency domain is obtained by applying Fourier transform to Eq. (6.15):

$$Y(j\omega) = H(j\omega) X(j\omega) \quad (6.16)$$

where $Y(j\omega)$ is the Fourier transform of output sequence:

$$Y(j\omega) = \sum_{n=-\infty}^{\infty} y[n] e^{-jn\omega} \quad (6.17)$$

$X(j\omega)$ is the Fourier transform of input sequence:

$$X(j\omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-jn\omega} \quad (6.18)$$

$H(j\omega)$ is the Fourier transform of impulse response $h[n]$:

$$H(j\omega) = \sum_{n=-\infty}^{\infty} h[n]e^{-jn\omega} \quad (6.19)$$

$H(j\omega)$ is also called the frequency response of the system, which represents the change of the amplitude and phase of the output sequence relative to the input sequence, it is generally a continuous function of ω . $H(j\omega)$ is a complex number, which can be written as:

$$H(j\omega) = |H(j\omega)|e^{j\varphi(\omega)} \quad (6.20)$$

$|H(j\omega)|$ is the amplitude response and $\varphi(\omega)$ is the phase response. It can be seen that the frequency response of the digital filter $H(j\omega)$ can change the weight of the frequency spectrum of the input sequence. This is the working principle of the digital filter. The filtering process is shown in Fig. 6.12, where Fig. 6.12a shows the square wave input and its spectrum, Figs. 6.6, 6.12b shows the impulse response of an ideal low-pass filter and its spectrum, Fig. 6.12c shows the output signal, which is the convolution of input signal and the impulse response in time domain, or the product of their spectra in frequency domain. It can be obviously seen from the amplitude response curve that signals with frequency $|\omega| \leq \omega_c$ can pass, and signals with other frequencies ($\omega_c \leq |\omega| \leq \pi$) have been attenuated. Comparing the waveforms of input $x[n]$ and output $y[n]$, we can see that the sharp change of the input signal has been smoothed, i.e. the system effectively filtered high frequency components in the input signal.

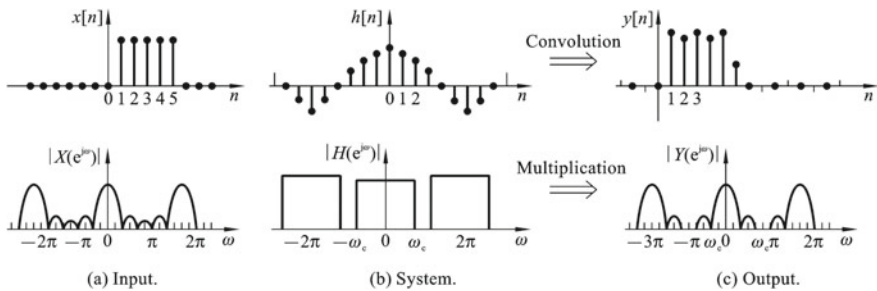


Fig. 6.12 Steps of digital filtering

6.2 Filtering in Frequency Domain

Broadly speaking, any kind of information transfer can be regarded as a kind of filtering. Its response is a function of input frequency, and can be described by a frequency domain function, i.e. the transfer function. Suppose the input signal is $x[n]$, the transfer function of the transfer device is $h[n]$, and the output signal is $y[n]$, then

$$y[n] = h[n] * x[n] \quad (6.21)$$

This is the time-domain filtering equation, and the filtering result is the convolution of the input signal $x[n]$ and the impulse response function $h[n]$ of the filter. According to the time domain convolution theory, it corresponds to the multiplication in frequency domain:

$$Y(f) = H(f)X(f) \quad (6.22)$$

This is the frequency domain filtering equation. The frequency spectrum of the output signal is equal to the product of the spectrum of input signal $X(f)$ and the spectrum of the impulse response function $H(f)$. Apply inverse Fourier transform, we get:

$$y[n] = F^{-1}[Y(f)] \quad (6.23)$$

Thus, the filtering result is:

$$y[n] = F^{-1}(X(f) \cdot H(f)) = F^{-1}(F(x[n]) \cdot H(f)) \quad (6.24)$$

The above formula is the equation for filtering the input $x[n]$ using the frequency domain filtering function $H(f)$, and the process is shown in Fig. 6.13.

Filters are generally divided into low-pass, high-pass, band-pass, and band-stop filters. Figure 6.14 shows the amplitude responses of the four filters.

Figure 6.14a shows the amplitude response of a low-pass filter, its transfer function is:

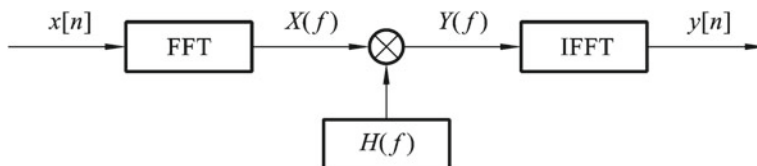
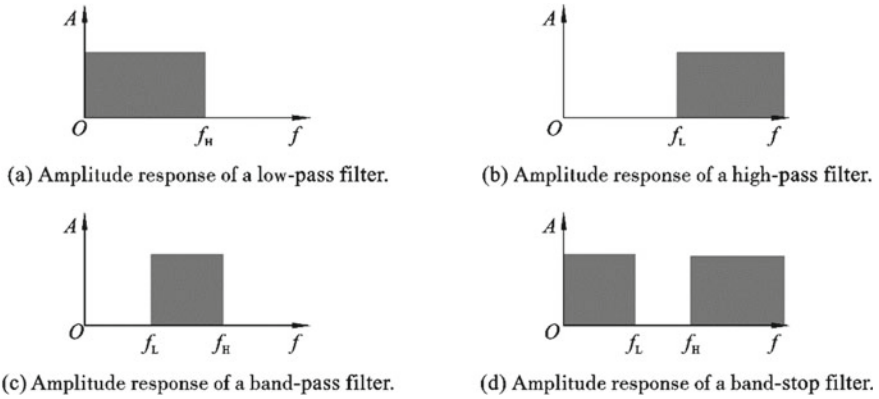


Fig. 6.13 Frequency domain filtering process

**Fig. 6.14** Amplitude responses of filters

$$H(f) = \begin{cases} 1 & f < f_H \\ 0 & \text{otherwise} \end{cases} \quad (6.25)$$

It allows low-frequency or DC components in the signal to pass, while suppresses high-frequency components or interferences and noises. The frequency components between 0 and f_H are passed without attenuation, while the rest of the frequency components are all attenuated.

Figure 6.14b shows the amplitude response of a high-pass filter, its transfer function is:

$$H(f) = \begin{cases} 1 & f > f_L \\ 0 & \text{otherwise} \end{cases} \quad (6.26)$$

In contrast to low-pass filter, it allows high-frequency components to pass while suppresses low-frequency and DC components. The frequency components higher than f_L are passed without attenuation, while other frequency components are attenuated.

Figure 6.14c shows the amplitude response of a band-pass filter, its transfer function is:

$$H(f) = \begin{cases} 1 & f_L < f < f_H \\ 0 & \text{otherwise} \end{cases} \quad (6.27)$$

It allows frequency components within a certain band (f_L, f_H) to pass, and suppresses frequencies below or above the frequency band.

Figure 6.14d shows the amplitude response of a band-stop filter, its transfer function is:

$$H(f) = \begin{cases} 1 & f < f_L \\ 1 & f > f_H \\ 0 & \text{otherwise} \end{cases} \quad (6.28)$$

It suppresses frequency components within a certain band (f_L, f_H) and allows other frequencies to pass. It is also called a notch filter.

As we have mentioned previously, low-pass filter and high-pass filter are the two most basic forms of filters. Band-pass filter is a serial connection of low-pass filter and high-pass filter:

$$H(f) = H_1(f)H_2(f)$$

Band-stop filter is low-pass filter and high-pass filter connected in parallel:

$$H(f) = H_1(f) + H_2(f)$$

For a practical filter, there is a transition zone between the pass band and the stop band. In the transition zone, signals will be attenuated to different degrees. This transition band is not desired by the filter, but it is also inevitable

Example 6.2: Use different types of filters to process the following signal in MATLAB

$$x(t) = \sin(2\pi \cdot 50t) + \sin(2\pi \cdot 300t) + \sin(2\pi \cdot 500t)$$

The signal has three frequency components: 50 Hz, 300 Hz and 500 Hz. Its time domain waveform and frequency domain spectrum are shown in Fig. 6.15. In the following MATLAB code, we design a low-pass filter with cut-off frequency of 150 Hz, multiply the spectrum of input signal $X(f)$ by the frequency response of the designed filter $H(f)$, and implement the inverse Fourier transform. Then the filtered signal can be obtained, and it is shown in Fig. 6.15c.

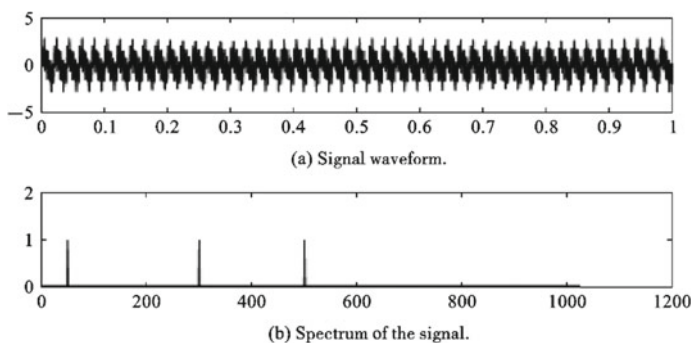
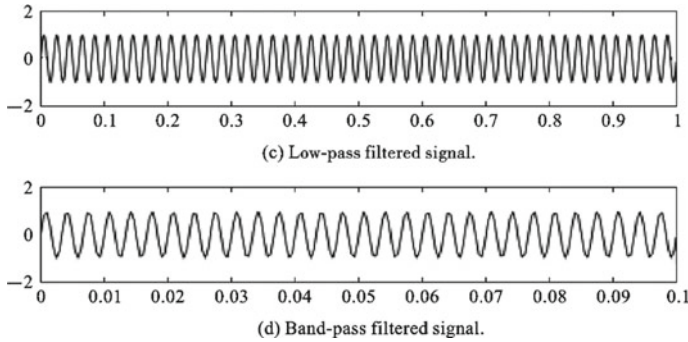


Fig. 6.15 Low-pass and band-pass filtering of a signal with multi-frequency

**Fig. 6.15** (continued)

```

Fs = 2048; dt=1.0/Fs; T =1;
N=T/dt; t=[0:N-1]/N;
x =sin(2*pi*50*t)+sin(2*pi*300*t) + sin(2*pi*500*t);
subplot(4,1,1); plot(t,x);
axis([0, 0.1, -2,2]);
X=fft(x,N);
P =2*abs(X)/N;
f=linspace(0,Fs/2,N/2);
subplot(4,1,2); plot(f,P(1:N/2));
df=Fs/N;
Fh=150; kh=floor(Fh/df);
H=ones(1,N);
for k=kh:N-kh+1
    H(k)=0;
end
Y=X.*H;
y = ifft(Y);
subplot(4,1,3); plot(t,y);
axis([0, 0.1, -2,2]);

```

We can also design a band-pass filter with cut-off frequency of 200–400 Hz. The filtered signal is shown in Fig. 6.15d. MATLAB code for the band-pass filtering is:

```

Fs = 2048; dt=1.0/Fs; T=1;
N=T/dt; t=[0:N-1]/N;
x =sin(2*pi*50*t)+sin(2*pi*300*t) + sin(2*pi*500*t);
subplot(4,1,1); plot(t,x);axis([0, 0.1, -2,2]);
X=fft(x,N);    P =2*abs(X)/N;
f=linspace(0,Fs/2,N/2); subplot(4,1,2);
plot(f,P(1:N/2)); df=Fs/N;
Fl=200;    kl=floor(Fl/df);
Fh=400; kh=floor(Fh/df);
H=ones(1,N);
for k=1:kl
    H(k)=0;
end
for k=N-kl+1:N
    H(k)=0;
end
for k=kh:N-kh+1
    H(k)=0;
end
Y=X.*H; y = ifft(Y);
subplot(4,1,4); plot(t,y);
axis([0, 0.1, -2,2]);

```

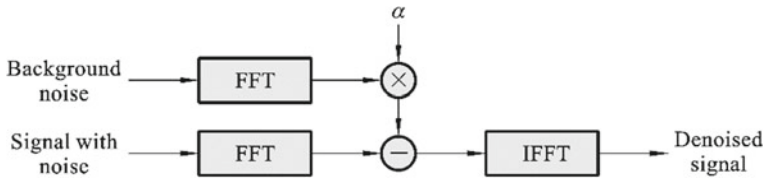


Fig. 6.16 Denoising by subtraction of noise spectrum

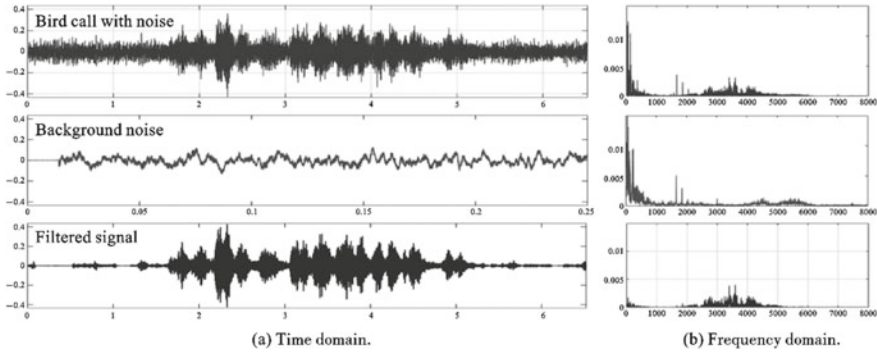


Fig. 6.17 Bird noise processing results

Example 6.3: Denoising by spectral subtraction

In the presence of stable background noise, spectral subtraction can be used to filter the background noise in the signal. Firstly, pure background noise is recorded and its spectrum is obtained by Fourier transform. Then the spectrum is multiplied by a correction coefficient α . After the measurement signal is recorded, we perform Fourier transform on the measurement signal, and subtract it by the spectrum of noise. Finally, we do the inverse Fourier transform to get the denoised signal. The process is shown in Fig. 6.16.

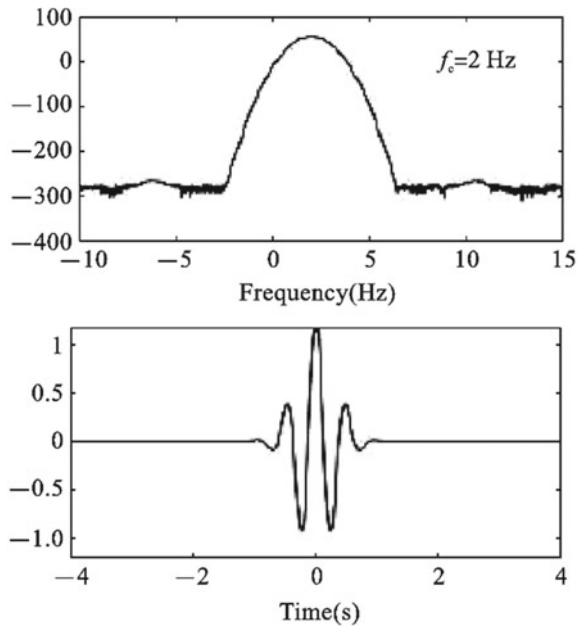
Figure 6.17 shows a fragment of bird sound with noise. Through the above processing, the background noise in signal can be eliminated.

6.3 Time Domain Filtering and Z-transform

6.3.1 Time Domain Filtering

If we regard the filter as a system, the Fourier transform of the transfer function $h[n]$ is $H(f)$. The input signal with noise is $x[n]$, and the output signal after denoising is $y[n]$. The output signal is equal to the convolution of the input signal and the impulse response function of the filter $y[n] = x[n] * h[n]$.

Fig. 6.18 Morlet wavelet band-pass filter



Example 6.4 Morlet wavelet band-pass filtering of a signal with four frequency components

$$x(t) = \sin(2\pi \cdot 20t) + \sin(2\pi \cdot 60t) + \sin(2\pi \cdot 120t) + \sin(2\pi \cdot 200t)$$

A Morlet wavelet band-pass filter is used for the filtering. The waveform and amplitude response of the filter is shown in Fig. 6.18.

The time domain function of the filter is:

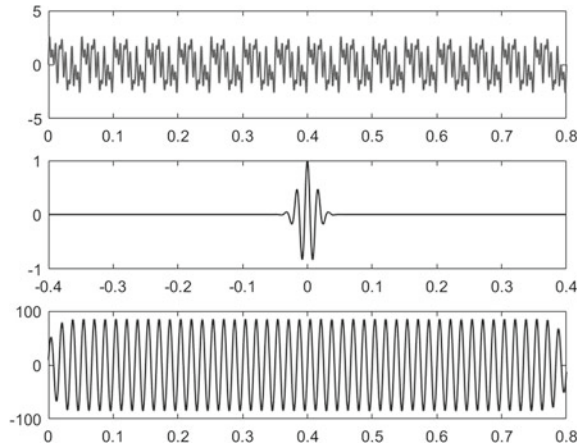
$$h(t) = e^{-(t/c)^2} \cos(2\pi f_c t)$$

And its transfer function is:

$$H(f) = e^{-(cf)^2} \delta(f - f_c)$$

The filtered signal can be calculated by the convolution $x[n]*h[n]$. The waveform and spectrum of the filtered signal are shown in Fig. 6.19. The MATLAB code for the filtering is:

Fig. 6.19 Time domain and frequency domain results after wavelet filtering



```

Fs = 5120;  dt=1.0/Fs;

N=4096; T=dt*N; t0=linspace(0,T,N);

x=sin(2*pi*20*t0)+sin(2*pi*60*t0)+sin(2*pi*120*t0)+sin(2*pi*200*t0);

subplot(3,1,1);

plot(t0,x,'linewidth',1);

%Morlet Wavelet Filter

Fc=60;

t1=linspace(-T/2,T/2,N);

f0=5/(2*pi); nn=Fc/f0;

x1=cos(2*pi*nn*f0*t1);

x2=exp(-nn*nn*t1.*t1/2);

wt=x1.*x2;

subplot(3,1,2);

plot(t1,wt,'r','linewidth',1);

y = conv(x,wt);

N1=length(y);

y=y(N/2:N+N/2-1);

subplot(3,1,3);

plot(t0,y,'r','linewidth',1);

```

6.3.2 Z-transform

Z-transform can be regarded as discretized Laplace transform and Fourier transform. It is a mathematical transformation performed on discrete sequences and is often used to find the solution of linear time-invariant difference equations. Z-transform has become an important tool for analyzing discrete linear time-invariant systems, and has a wide range of applications in digital signal processing, computer control systems and other fields.

Z-transform can transform time domain signal (i.e. discrete time sequence) into the complex frequency domain. Its importance in discrete-time signal processing is like the Laplace transform in continuous-time signal processing. In Z-transform, the time-domain mathematical model of discrete linear time-invariant system is transformed into algebraic equations in the Z domain, which simplifies the analysis of discrete systems. We can also use system functions to analyze the time domain characteristics, frequency response and stability of the system.

Z-transform has many important properties, such as linearity, time-shifting, differentiation, sequential convolution, complex convolution. These properties all play an important role in solving signal processing problems. Among them, the property of convolution is the most important one. Since the task of signal processing is to output the required signal sequence after the input signal sequence is processed by a certain system, it is important to find a way to compute output signal by the input signal and the system characteristic. Through theoretical analysis, it can be known that if the computation is to be performed in time domain, tedious convolution operations must be done. Using the convolution property of the Z-transform, the process can be greatly simplified. As long as the Z-transforms of the input signal sequence and the system impulse response sequence are obtained, and the inverse transform of the product of the two is obtained, the output signal sequence can be obtained. The inverse transform here is the inverse Z-transform.

Currently, there are already tables for the result of Z-transform and Laplace transform. For common signal sequences, the Z-transform result can be directly found on the table. Correspondingly, if the Z-transform result is known, the original signal sequence can also be found by the table.

In the theoretical research of Z-transform, W. Hurewicz took the first step in 1947. He first introduced a transform to process discrete sequences. Later John Ragazzini and Lotfi Zadeh greatly simplified the calculation steps and named it Z-transform in 1952. Since the Z-transform can only reflect the law of the impulse system at the sampling point, Eliahu Jury proposed the modified or advanced Z-transform in 1956.

For discrete sequences, Z-transform is actually an operation of shifting. For a Z-transform system with the transfer function of

$$H(z) = z^{-1}$$

the transform result $y[n]$ of input $x[n]$ is actually $y[n] = x[n-1]$. For another Z-transform with transfer function of:

$$H(z) = z^2$$

the transform result $y[n]$ of input $x[n]$ is $y[n] = x[n + 2]$. Therefore, for a filter with transfer function of $H(z)$, it is easy to directly obtain the filtering equation in time domain.

Example 6.5: Z-transform in MATLAB for half-band filter with the following transfer function

$$H(z) = 0.006614 + 0.004007z^{-1} - 0.077517z^{-2} + 0.068953z^{-3} + 0.496251z^{-4} \\ + 0.496251z^{-5} + 0.068953z^{-6} - 0.077517z^{-7} + 0.004007z^{-8} + 0.006614z^{-9}$$

Its amplitude response is show in Figs. 6.20 and 6.21.

The MATLAB code for using the filter is listed below:

Fig. 6.20 Half-band filter

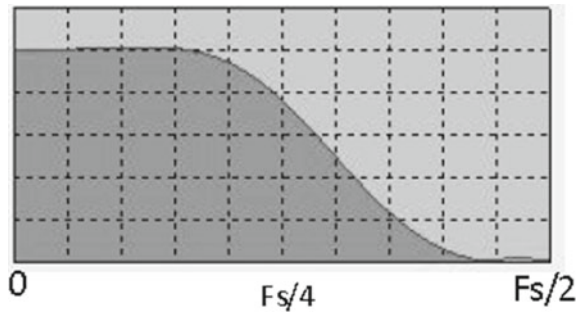
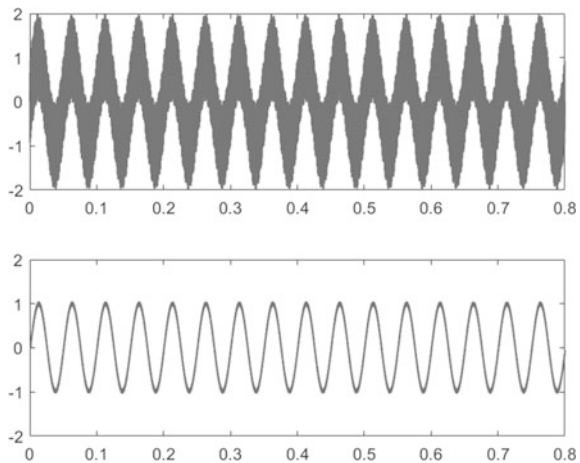


Fig. 6.21 Z-transform by MATLAB



```

Fs = 5120; dt=1.0/Fs;
N=4096; T=dt*N; t0=linspace(0,T,N);
x=sin(2*pi*20*t0)+sin(2*pi*2000*t0);
subplot(2,1,1); plot(t0,x,'linewidth',1);
y(1)=0;y(2)=0;y(3)=0;
y(4)=0;y(5)=0;y(6)=0;
y(7)=0;y(8)=0;y(9)=0;
for k=10:N
y(k)=0.006614+0.004007*x(k-1)-0.077517*x(k-2)+0.068953*x(k-3)+0.496251*x(k-4)+0.496251
*x(k-5)+0.068953*x(k-6)-0.077517*x(k-7)+0.004007*x(k-8)+0.006614*x(k-9);
end
subplot(2,1,2);
plot(t0,y,'linewidth',1);

```

6.3.3 Bilateral Z-transform

Bilateral Z-transform is also called two-sided Z-transform. It is defined as:

$$X(Z) = Z\{x[n]\} = \sum_{n=-\infty}^{\infty} x[n]Z^{-n}, \quad Z \in R(x) \quad (6.29)$$

$R(x)$ is called the region of convergence (ROC) of $X(z)$.

Unilateral Z-transform is also called single-sided Z-transform. It is defined as:

$$X(Z) = Z\{x[n]\} = \sum_{n=0}^{\infty} x[n]Z^{-n}, \quad Z \in R(x) \quad (6.30)$$

where $R(x)$ is the region of convergence

6.4 Finite Impulse Response (FIR) Filter

If the filter is in the zero state, according to Z-transform, the convolution of Eq. (6.22) can be transformed into:

$$Y(z) = H(z)X(z) \quad (6.31)$$

where $Y(z)$ is Z-transform of the zero-state response of $y[n]$, $X(z)$ is the Z-transform of excitation signal $x[n]$, $H(z)$ is the Z-transform of the unit impulse response $h[n]$ and is called system function. From Eq. (6.30), the following equation can be derived:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{r=0}^M b_r z^{-r}}{\sum_{k=0}^N a_k z^{-k}} \quad (6.32)$$

If the system function $H(z)$ is a rational function. Then the numerator and denominator of Eq. (6.32) can be decomposed into factors:

$$H(z) = G \cdot \frac{\prod_{r=1}^M (1 - z_r z^{-1})}{\prod_{k=0}^N (1 - p_k z^{-1})} \quad (6.33)$$

where z_r ($r = 1, 2, \dots, M$) are system zeros, p_k ($k = 1, 2, \dots, N$) are system poles. Obviously, the zeros and poles are determined by the numerator and denominator of $H(z)$, and the distribution of zeros and poles can determine the property of the system and the property of the unit impulse response.

Digital filter is a discrete linear time-invariant system, and its system function $H(z)$ is a rational function of z^{-1} . Therefore, Eq. (6.32) can be written as:

$$H(z) = \frac{\sum_{r=0}^M b_r z^{-r}}{1 + \sum_{k=1}^N a_k z^{-k}} \quad (6.34)$$

It can be seen that, if $a_k = 0$ ($k = 1, 2, \dots, N$), then $H(z)$ is a polynomial of z^{-1} :

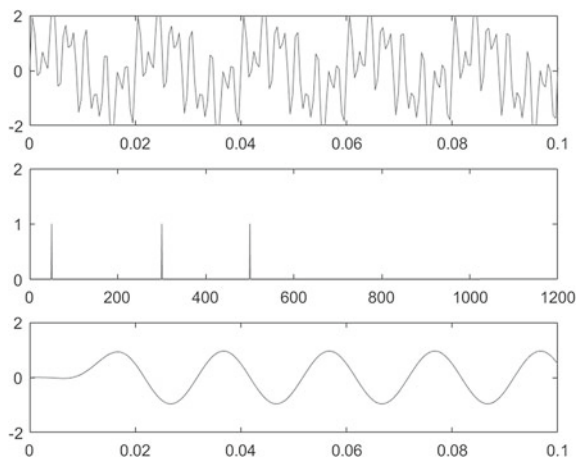
$$H(z) = \sum_{r=0}^M b_r z^{-r}$$

Namely, the corresponding unit impulse response $h[n]$ has a finite length, thus the corresponding filter is called a finite impulse response (FIR) filter.

Example 6.6: FIR Filter Design in MATLAB

The following MATLAB code is to design a low-pass FIR filter to remove the 300 Hz and 500 Hz frequency components in a signal. The result is shown in Fig. 6.22.

Fig. 6.22 FIR filtering in MATLAB



```

Fs = 2048; dt=1.0/Fs;

T = 1; N=T/dt; t=[0:N-1]/N;

x1 =sin(2*pi*50*t)+sin(2*pi*300*t) +sin(2*pi*500*t);

subplot(3,1,1); plot(t,x1);

axis([0, 0.1, -2,2]);P=fft(x1,N);

Pyy =2*sqrt(P.* conj(P))/N;

f=linspace(0,Fs/2,N/2);

subplot(3,1,2); plot(f,Pyy(1:N/2));

b = fir1(48,0.1);

x2= filter(b,1,x1);

subplot(3,1,3); plot(t,x2);

axis([0, 0.1, -2,2]);

```

6.5 Infinite Impulse Response (IIR) Filter

If $a_k \neq 0$ in Eq. (36), then the unit impulse response $h[n]$ is infinite, and the corresponding filter is called an infinite impulse response (IIR) filter.

According to the system function of digital filter, the digital convolution integral equation can be used to filter the digitized measurement signal $x[n]$:

$$y[n] - a_1 y[n-1] - \dots - a_{N_1} y[n-N_1] = b_0 x[n] - b_1 x[n-1] - \dots - b_{N_2} x[n-N_2], n = 0, 1, \dots, M \quad (6.35)$$

where $y[n]$ is the filtered signal, M is the length of sampling signal.

Whether it is a FIR or IIR digital filter, the filtering process is very simple. Using a double loop statement structure, the digital filtering process of Eq. (6.35) can be realized by operations of adding, subtracting, multiplying. The difficulty posed is how to quickly construct the system function $H(z)$ of the digital filter according to the needs, and obtain the filter coefficients.

FIR filter design methods include window function method, frequency sampling method, optimization method, etc.; IIR filter design methods include bilinear transform method, digital Butterworth filter, digital Chebyshev filter and so on. The design of digital filters involves a lot of mathematical knowledge, which is difficult to explain here. However, there are many software toolkits that have provided subroutines for the filter design. Users can call these subroutines directly to design a filter easily.

Example 6.7: Design of IIR Filter in MATLAB

The following MATLAB code is to design a low-pass Butterworth filter to remove the 300 Hz and 500 Hz frequency components in a signal. You can also try to comment the code with “butter” function and uncomment the “ellip” function to use an elliptic filter (Fig. 6.23).

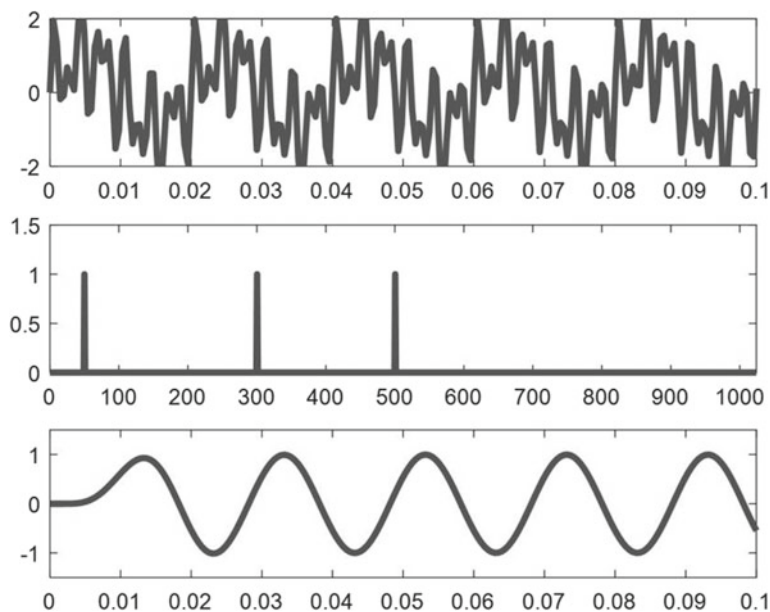


Fig. 6.23 IIR filtering in MATLAB

```

Fs = 2048; dt=1.0/Fs;

T =1; N=T/dt; t=[0:N-1]/N;

x1 =sin(2*pi*50*t)+sin(2*pi*300*t)+sin(2*pi*500*t);

subplot(3,1,1); plot(t,x1,'linewidth',3);

axis([0, 0.1, -2,2]);

P=fft(x1,N);

Pyy =2*sqrt(P.* conj(P))/N;

f=linspace(0,Fs/2,N/2);

subplot(3,1,2); plot(f,Pyy(1:N/2),'linewidth',3);

axis([0,1024, 0,1.5]);

[b,a] = butter(8,0.1,'low');

%[b,a] =ellip(8,1,60,0.1,'low');

x2= filter(b,a,x1);

subplot(3,1,3); plot(t,x2,'linewidth',3);

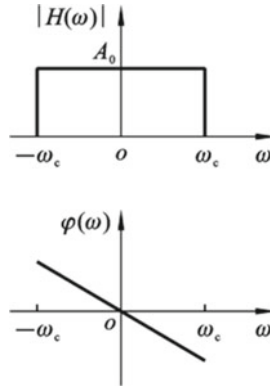
axis([0, 0.1, -1.5,1.5]);

```

6.6 Exercise

- 6-1 What is signal filtering, and under what conditions can it work well?
- 6-2 Turn on and off the noise cancellation function of your cell phone, talk to your friends in a noisy environment, and compare their differences.
- 6-3 For a low-pass filter with the following transfer function

$$H(\omega) = \begin{cases} A_0 e^{-j\omega\tau_0} & -\omega_c < \omega < \omega_c \\ 0 & \text{otherwise} \end{cases}$$



Exercise Fig. 6.1

when a δ function passed it, try to:

- (1) find the time domain waveform of the output signal
- (2) find the output spectrum
- (3) explain why the filter is a non-causal system

6-4 For an ideal low-pass filter

$$H(\omega) = \begin{cases} A_0 e^{-j\omega\tau_0} & -\omega_c < \omega < \omega_c \\ 0 & \text{otherwise} \end{cases}$$

when a unit step function passed it, try to:

- (1) find the time domain waveform of the output signal
- (2) find the output spectrum
- (3) relationship between bandwidth of filter and rise time
- (4) explain the Gibbs phenomenon that appeared

6-5 For an ideal filter with transfer function of $H(j\omega)$ and unit step response $y_M(t)$

$$H(j\omega) = e^{-j\omega\tau_0} \quad -\omega_c < \omega < \omega_c$$

$$y_M(t) = \frac{1}{2} + \frac{1}{\pi} \text{si}[\omega_c(t - t_0)]$$

when a rectangular pulse

$$x_r(t) = u(t) - u(t - \tau)$$

passed, try to find:

- (1) time domain waveform
- (2) frequency spectrum
- (3) how the waveform changes with the change of bandwidth ω_c .

6-6 For a band-pass filter with transfer function

$$H(j\omega) = \frac{1}{1 + j(\omega - 100)}$$

when an amplitude modulated signal

$$x_A(t) = (1 + \cos t) \cos 100t$$

passed the filter, try to find the waveform of the output signal and its spectrum.

6-7 An ideal low-pass filter has the following characteristics: $|H(\omega)| = 1$, $\varphi(\omega) = 4 \times 10^{-6} \omega$, $\omega_c = 6 \times 10^5$ rad/s. If a step function of 12 V is used as input, try to find:

- (1) amplitude of the output signal 6 μ s after the excitation
- (2) the time it takes for the output signal amplitude to reach 6 V.

6-8 What is Z-transform and why do we need it in time domain filtering?

6-9 Calculate the step response of the first-order system using the Z-transform.

Chapter 7

Principles of Sensors



7.1 Overview of Sensor Technology

7.1.1 Definition of Sensor

A sensor is a device that receives one form of information (physical, chemical, biological, etc.) and converts it into another form of information (usually electrical and optical) according to a certain rule. Nowadays, electrical signals can be processed and transmitted more easily. Therefore, a sensor can be narrowly defined as a device that converts non-electrical signals into electrical signals.

Sensor is the first unit of a measuring and control system, it is used to sense the stimulus and changes in the environment. The role of a sensor is very similar to the role of the sense organs of the human body. A comparison between the measuring and control system and the human body is shown in Fig. 7.1. The five sense organs (eyes, ears, nose, tongue and skin) are specialized organs in human body that help us perceive the world around us, the sense organs can sense the stimulus such as light and sound, and send electrical signals to our brain via the nervous system. Then, our limbs can react according to the stimulus after the analyses in brain. In a measuring and control system, the computer is acting as the brain to analyze the signal acquired by sensors and sends a signal to control the actuators.

Since the role of sensors is similar to the role of sense organs, we can make an analogy between them in Table 7.1. The functions of some sensors have already surpassed our sense organs, e.g. the CCD has better sensitivity than our eyes and infrared sensor have larger detectable frequency range than our eyes. Yet, there are still some sensors that are not as good as sense organs such as the ion sensors for taste.

As we will see in the following sections, sensors have been widely used in our daily life and many industrial fields to obtain accurate information from outside world. Therefore, sensor technology, communication technology and computer technology are considered as the three pillars of modern information industry.

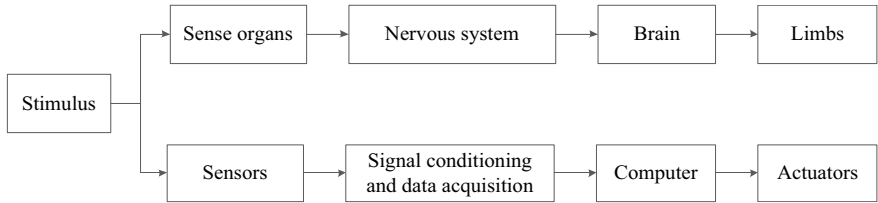


Fig. 7.1 Comparison between the measuring and control system and the human body

Table 7.1 Analogy between sensors and sense organs

Sense	Sense organs	Sensors
Sight	Eyes	CCD, CMOS, infrared sensor
Hearing	Ears	Acoustic sensor, ultrasonic sensor
Smell	Nose	Gas sensor
Taste	Tongue	Ion sensor
Touch	Skin	Pressure sensor, temperature sensor

7.1.2 Composition of Sensors

Usually, a sensor is composed of a conversion structure and a sensing element as shown in Fig. 7.2. When the physical quantity to be measured is difficult to be converted into electrical signals directly, a conversion structure is used convert the physical quantity into another intermediate physical quantity. Then, the sensing element converts the intermediate physical quantity into electrical signal.

Example 7.1: Digital Hanging Scale

Digital hanging scale (Fig. 7.3) is a simple example of sensor. The physical quantity to be measured by the scale is weight. Since it is difficult to convert weight into electrical signal directly, the spring is used as the conversion structure to convert the weight into displacement according to Hook’s law. Then, the moving end of the spring is connected to a rheostat. The rheostat is acted as the sensing element to convert the displacement into changes of electrical quantities.

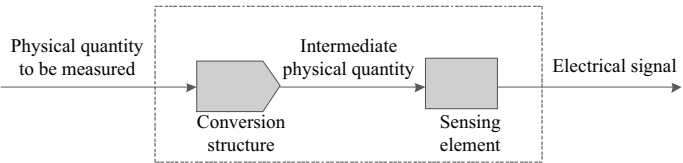


Fig. 7.2 Composition of a sensor

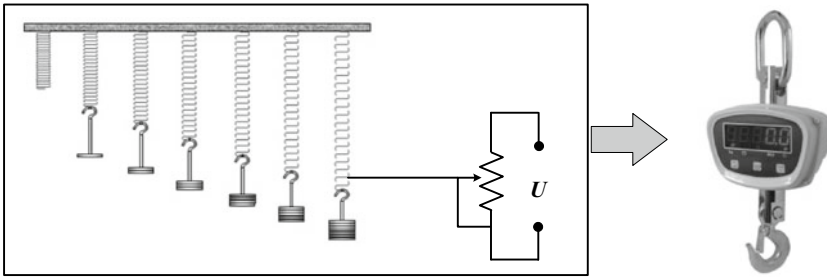


Fig. 7.3 Schematic of a digital hanging scale

Example 7.2: Digital Kitchen Scale

Digital kitchen scale is also an example of sensor. The physical quantity to be measured is still the weight. The conversion structure for the kitchen scale is the cantilever, which converts the weight into the strain of the cantilever. Then, resistive strain gauge is used as the sensing element to convert the strain into electrical signal.

7.1.3 Classifications of Sensors

Various sensors can be applied to measure different quantities such as length, area, volume, displacement, velocity, acceleration, flow speed, vibration, pressure, force, torque, weight, temperature, humidity, luminance, magnetic field intensity, sound, particle concentration, etc. They are based on different working principles such as the electromagnetic induction, piezoelectric effect, photovoltaic effect, thermoelectric effect, etc. Therefore, there are many types of sensors and many ways to classify them. Only some commonly used classification methods are introduced in this book.

1. Classification according to the application (the quantity to be measured)

The simplest way to classify and name a sensor is according to the quantity to be measured. For example, a sensor used to measure displacement is called a displacement sensor. Table 7.2 lists some of the typical sensors named after the physical quantity to be measured.

2. Classification according to the working principle

Sensors convert physical quantities into electrical signal according to different principles. We also use the working principle to classify and name a sensor. For example, if the sensor is based on Hall effect, we call it Hall sensor. And if a sensor converts the physical quantity to be measured into inductance, we call it inductive sensor.

3. Classification according to the type of energy supply

According to the type of energy supply, sensors can be classified as energy conversion type and energy control type. For the energy conversion type, the energy is self-supplied. The sensor energy comes directly from the measured physical quantity. For

Table 7.2 Sensors classified according to the quantity to be measured

Sensors	Quantity to be measured
Mechanical	Length, thickness, displacement, velocity, acceleration, rotational angle, rotational speed, mass, weight, force, pressure, torque
Acoustic	Sound pressure, noise
Magnetic	Magnetic flux, magnetic field intensity
Thermal	Temperature, heat, specific heat capacity
Optic	Brightness, color

example, the thermocouple directly converts the heat energy into electrical energy and the piezoelectric transducer converts the force or pressure into electrical energy. While for the energy control type, the energy is provided by an external power supply. The measured quantity only controls the output energy. For example, for the strain gauge, a power supply must be connected to it. The energy of the output signal comes from the power supply. The measured strain only controls the amount of energy in the output.

4. Classification according to the characteristics of sensing element

Sensors can be classified as physical and structural types according to the characteristics of sensing element. If signal conversion is made by the change of physical characteristics, the sensor is classified as a physical sensor. For example, the mercury thermometer belongs to the physical type because the temperature changes the volume, which is a physical characteristic, of the mercury. The magnetoresistor also belongs to the physical type because the magnetic field changes the resistance of the sensing element. For the structural type, the signal conversion is made by the change of structural parameters. For example, when we use the capacitive sensor to measure sound intensity, the sound (mechanical wave) changes the structural characteristic (distance between two plates) of the capacitor.

7.2 Resistive Sensors

A resistive sensor is a sensor that converts the measured quantity into the change of resistance. According to the working principle, it can be divided into: potentiometer, resistive strain gauge and sensitive type (thermosensitive, photosensitive, gas sensitive, magneto sensitive, etc.). Resistive sensor has a wide range of applications because resistor is the simplest electronic component and the measurement of resistance is simple, accurate and has a large dynamic range. A multimeter can measure resistance from 10 Ω to 10 MΩ, and the change of resistance can be easily converted to voltage or current by Ohm’s law $U = I \times R$.

7.2.1 Potentiometer

Potentiometer is a three-terminal resistor with sliding or rotating contacts to form an adjustable voltage divider. It has been widely used to measure displacement and angle. The resistance of a wire can be represented by the following equation:

$$R = \rho \frac{l}{S} \quad (7.1)$$

where ρ is the resistivity, l is the length of the wire and S is the cross-sectional area of the wire. When the wire is made of the same material and the cross-sectional area is kept constant for the whole wire, the resistance of a wire is proportional to its length.

An equivalent circuit of the potentiometer is shown in Fig. 7.4. The resistance between Terminal A and B and the resistance between B and C are respectively:

$$R_{AB} = kL \quad (7.2)$$

$$R_{BC} = kx \quad (7.3)$$

where k is the resistance per unit length, L is the total length of the potentiometer and x is the displacement of terminal C. Terminal A and B are connected to a DC power supply and the voltage across terminal B and C is used as the output. Since R_{AB} and R_{BC} forms a voltage divider, the following expression can be easily obtained:

$$U_{\text{out}} = \frac{R_{BC}}{R_{AB}} U = \frac{U}{L} x \quad (7.4)$$

where U is the voltage of the power supply. It can be seen from Eq. (7.4) that the output voltage of a potentiometer is proportional to the displacement of the moving terminal.

If the output of the potentiometer is further connected to other instruments or it is used to drive an actuator, the load resistance must be considered and the circuit is

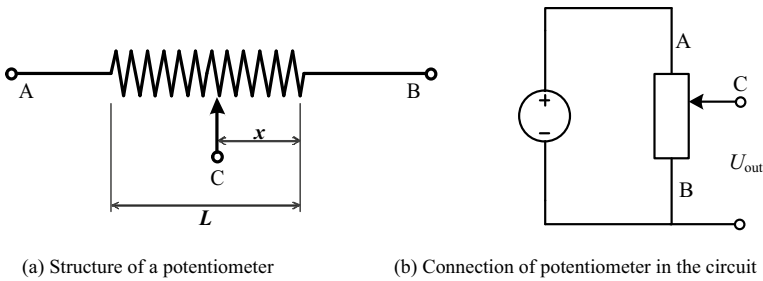
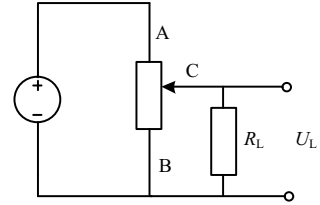


Fig. 7.4 Equivalent circuit of the potentiometer

Fig. 7.5 Equivalent circuit of potentiometer with load



shown in Fig. 7.5. The voltage applied on the load is:

$$U_L = \frac{R_{BC} \parallel R_L}{R_{BC} \parallel R_L + R_{AC}} U = \frac{U}{\frac{L}{x} + \frac{R_{AB}}{R_L} (1 - \frac{x}{L})} \quad (7.5)$$

where R_L is the load resistance. From Eq. (7.5), we can find that the output voltage is no longer proportional to the displacement. A comparison between the output voltages of the potentiometer in open circuit and with load is made in Fig. 7.6. The non-linear relationship will cause errors in the measurement. It can be noticed that, if the load resistance in Eq. (7.5) goes to infinity, the output voltage will be proportional to the displacement again. Therefore, when we are using the potentiometer, the measuring instrument should have a large internal resistance.

When selecting a potentiometer for our application, usually there six parameters to be considered:

- (1) Linearity, which is a parameter used to describe the goodness of the linear relationship between the output voltage and input displacement or angle;
- (2) Resolution, which describes the smallest measurable interval of displacement or angle;
- (3) Deviation of the total resistance;
- (4) Measuring range, which describes the maximum value of displacement or angle that can be measured;
- (5) Temperature coefficient, which describes if the resistance of the potentiometer changes with temperature;
- (6) Lifespan.

Fig. 7.6 Output voltages of the potentiometer in open circuit and with load

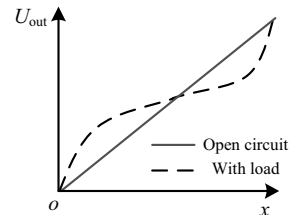
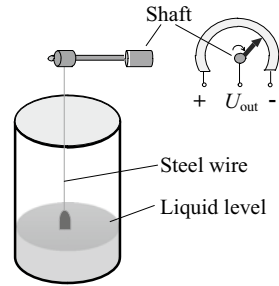


Fig. 7.7 Detection of gas reservation with potentiometer



Example 7.3: Detection of Gas Reservation

In factories, gas is an important source of fuel, so it is necessary to know how much gas is left in the tank. If the gas is not enough, the workers should add it in time. Due to the high pressure in the tank, the gas is in the liquid form. Figure 7.7 illustrates the schematics of measuring gas reservation with a potentiometer. A buoy is floating on the liquid surface and its vertical position changes with the liquid level. The buoy is connected to a shaft through a steel wire, so that movement of the buoy will rotate the shaft. The shaft is further connected to the moving terminal of a potentiometer, thus the rotation of the shaft will change the resistance of the potentiometer. If we connect the potentiometer in a circuit, the output voltage will change with the liquid level in the tank.

Example 7.4: Resistive Touchscreen

Touchscreen is an intuitive input and output device for modern instruments. Decades ago, resistive touchscreens were used in many smartphones such as Nokia N97 and Samsung Omnia. Nowadays, most smartphones use capacitive touchscreen while the resistive touchscreen is mainly used in the instruments in hospitals and factories due to its tolerance for liquids. A typical resistive touchscreen is consisted of two independent layers as shown in Fig. 7.8a, they have electrodes in x -axis and y -axis respectively. Each layer is made of grids of resistors as shown in Fig. 7.8b. Without touching, the two layers form two independent circuits as shown in Fig. 7.8d. If our finger touches the screen, the pressure will connect the two layers at the touching point and the circuits in the two layers are connected. To get the coordinate of the touching point, we need to read the x and y coordinates one by one. To read the x coordinate, we need to apply a drive voltage U_d to the $x+$ terminal and ground the $x-$ terminal, then read the voltage at $y+$ or $y-$ terminal. Since the circuit is a voltage divider, the voltage at $y+$ terminal is $U_{y+} = U_d x / W$, where x is the coordinate of the touching point and W is the total width of the touchscreen. Similarly, we can apply a DC voltage to the $y+$ and $y-$ terminals and read from the $x+$ terminal, then the output voltage will be proportional to the y coordinates: $U_{x+} = U_d y / H$, where H is the height of the touchscreen.

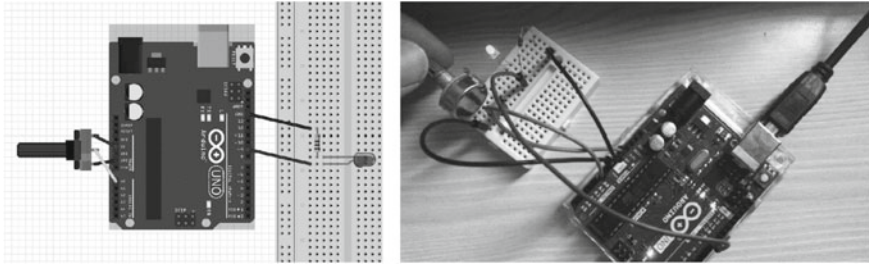


Fig. 7.9 Circuit connection for the potentiometer and LED

```

const int analogInPin = A0;

const int analogOutPin = 9;

int sensorValue = 0;

int outputValue = 0;

void setup() {
    Serial.begin(9600);
}

void loop() {
    sensorValue = analogRead(analogInPin); // read the analog in value
    outputValue = map(sensorValue, 0, 1023, 0, 255);
    analogWrite(analogOutPin, outputValue); // change the analog out value
    delay(100);
}

```

7.2.2 Resistive Strain Gauge

Resistive strain gauges are made of metal or semiconductor wires. The resistance of strain gauge changes under external strain due to the deformation. Strain gauge can be used to measure strain, force, velocity, acceleration, torque, etc. It has the advantages of small size, fast dynamic response, and high precision. It has been used in industries such as aerospace, marine, mechanics and architecture.

1. Metallic strain gauge

The resistance of a metal wire has already been displayed in Eq. (7.1). Take the derivative of the resistance, the following expression can be obtained:

$$\begin{aligned} dR &= \frac{\partial R}{\partial l} dl + \frac{\partial R}{\partial S} dS + \frac{\partial R}{\partial \rho} d\rho \\ &= \frac{\rho}{S} dl - \frac{\rho l}{S^2} dS + \frac{l}{S} d\rho \\ &= R \left(\frac{dl}{l} - \frac{dS}{S} + \frac{d\rho}{\rho} \right) \end{aligned} \quad (7.6)$$

where S is the cross-sectional area. Substituting $S = \pi r^2$ into Eq. (7.6), we get:

$$\frac{dR}{R} = \frac{dl}{l} - \frac{2dr}{r} + \frac{d\rho}{\rho} \quad (7.7)$$

where r is the radius of the wire. When a metal wire deforms, its volume is kept unchanged. Thus, when the wire elongates in the axial direction, the radius should reduce accordingly. The changes of length and radius are related by the Poisson's ratio ν as:

$$\frac{dr}{r} = -\nu \frac{dl}{l} \quad (7.8)$$

By definition, the ratio of the length variation to the original length dl/l is strain, and is usually denoted by ε . Besides, the change of resistivity is related to the axial stress:

$$\frac{d\rho}{\rho} = K_\pi \sigma = K_\pi E \varepsilon \quad (7.9)$$

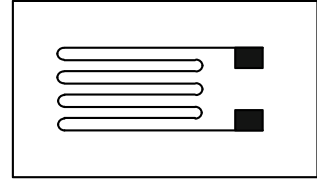
where K_π is piezoresistive coefficient, E is the Young's modulus, σ is the stress and ε is the strain. Substituting Eqs. (7.8) and (7.9) into Eq. (7.7), we can finally relate the change of resistance to strain as:

$$\frac{dR}{R} = \varepsilon + 2\nu\varepsilon + K_\pi E \varepsilon = (1 + 2\nu + K_\pi E) \varepsilon \quad (7.10)$$

The piezoresistive coefficient for metal is usually very small, thus the resistivity of metal is considered as a constant under stress. So Eq. (7.10) can be simplified as:

$$\frac{dR}{R} \approx (1 + 2\nu) \varepsilon \quad (7.11)$$

Fig. 7.10 Metallic strain gauge structures



Equation (7.11) indicates that the relative change of resistance is proportional to the strain and they have a linear relationship. The gauge factor, which indicates the sensitivity of the strain gauge, can be obtained from Eq. (7.11):

$$K_{MS} = 1 + 2\nu \quad (7.12)$$

where K_{MS} is the gauge factor. The Poisson's ratio for most metals is around 0.33. Therefore, the metallic strain gauge has an approximate gauge factor of 1.66.

A typical metallic strain gauge consists of metallic foils (or wires) in a zig-zag pattern of parallel lines as shown in Fig. 7.10. The metallic foils are supported by a flexible backing made of insulating materials. Usually, there is also a cover layer above the metallic foil to protect it from abrasion, corrosion and moisture environment. The objective of using parallel lines is to increase the change of resistance under the same strain. Although the parallel structure does not change the gauge factor, it increases the total resistance R in Eq. (7.11). Therefore, under the same strain ε , the change of resistance dR is increased.

The strain gauge shown in Fig. 7.10 has much larger sensitivity in the horizontal direction than in the vertical direction. In order to measure strains in different directions, we can use the structures shown in Fig. 7.11.

2. Semiconductor strain gauge

For the semiconductor strain gauge, the change of resistance is also governed by Eq. (7.10). The difference is that the semiconductor has a much larger piezoresistive coefficient, usually it is between $4 \times 10^{-10} \text{ m}^2/\text{N}$ to $8 \times 10^{-10} \text{ m}^2/\text{N}$. If we take the Young's modulus as $1.5 \times 10^{11} \text{ N/m}^2$, then $K_{\pi}E$ is approximately 60–120, which is much larger than $1 + 2\nu$. Therefore, the change of resistance of the semiconductor strain gauge is approximately expressed as:

Fig. 7.11 Metallic strain gauge with multi-directions

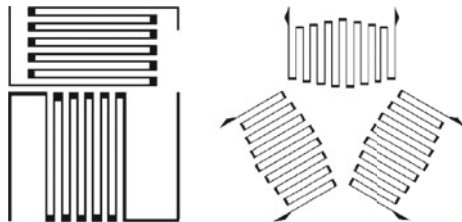
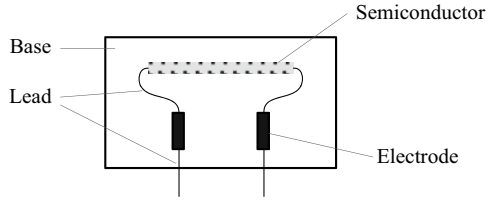


Fig. 7.12 Semiconductor strain gauge



$$\frac{dR}{R} = K_{\pi} E \varepsilon \quad (7.13)$$

The gauge factor is:

$$K_{SS} = K_{\pi} E \quad (7.14)$$

The main advantage of the semiconductor strain gauge is its high sensitivity, usually is around 100 times of the metallic strain gauge. In addition, it has quicker response and smaller size. However, it also has the disadvantages of temperature drift and bad linearity when measuring large strain.

Due to its large gauge factor, the semiconductor strain gauge usually only consists of single wafer or filament as shown in Fig. 7.12. The filament is bonded to an insulating base. Gold leads are used to connect the semiconductor to electrodes and external circuits.

3. Measurement circuits for strain gauge

As we can see from Eqs. (7.11) and (7.13), strain gauges can convert strain into the change of resistance. However, the change of resistance is usually very small and it is difficult to be measured directly. Thus, measurement circuits are needed to further convert the change of resistance into the change of voltage.

Wheatstone bridge is the most commonly used circuit for strain gauges due to its high sensitivity to small change of resistance. The DC Wheatstone bridge consists of four resistors as shown in Fig. 7.13. The output voltage of the bridge is:

$$U_{\text{out}} = \frac{R_4}{R_2 + R_4} U - \frac{R_3}{R_1 + R_3} U = \frac{R_1 R_4 - R_2 R_3}{(R_1 + R_3)(R_2 + R_4)} U \quad (7.15)$$

Fig. 7.13 Wheatstone bridge

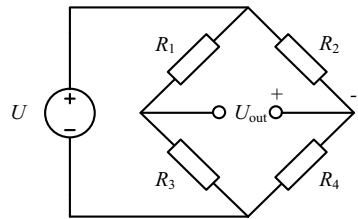
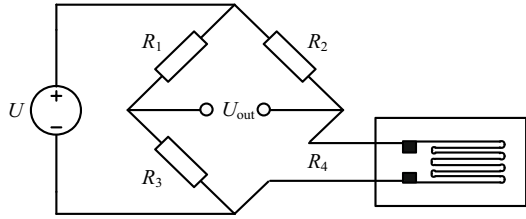


Fig. 7.14 Quarter bridge for strain gauge



The balance condition for Wheatstone bridge is:

$$\frac{R_1}{R_3} = \frac{R_2}{R_4} \quad (7.16)$$

Under the balance condition, the output voltage is zero. But when the resistance of one resistor changes, the bridge will output non-zero voltages.

When using Wheatstone bridge for strain gauges, we can replace one of the resistors in the bridge with a strain gauge to build a quarter bridge as shown in Fig. 7.14. To increase the sensitivity and linearity of the circuit, we need to set the initial condition as $R_1 = R_2 = R_3 = R_4 = R$. Without strain, the output voltage would be zero. And when the resistance of the strain gauge changes from R to $R + dR$ under the strain, the output voltage is:

$$U_{\text{out}} = \frac{R(R + dR) - RR}{(R + R)(R + R + dR)} U = \frac{dR}{4R + 2dR} U \approx \frac{U}{4} \frac{dR}{R} \quad (7.17)$$

By combining Eq. (7.17) with Eq. (7.11) or (7.13), we can find that the output voltage is proportional to the strain as:

$$U_{\text{out}} = \frac{U}{4} K_{\text{GF}} \varepsilon \quad (7.18)$$

where K_{GF} is the gauge factor for metallic or semiconductor strain gauge, namely K_{MS} or K_{SS} .

When the strain to be measured is small, multiple strain gauges can be used together to increase the sensitivity as shown in Fig. 7.15. In the figure, strain gauge 1# (R_2) is under tension and strain gauge 2# (R_4) is under compression. We can connect them to the Wheatstone bridge to form the half bridge shown in Fig. 7.16a. The output voltage can be easily calculated as:

$$U_{\text{out}} = \frac{U}{2} \frac{dR}{R} \quad (7.19)$$

By comparing with the output voltage of quarter bridge in Eq. (7.17), we can find that the sensitivity of the half bridge is doubled. The sensitivity can be further

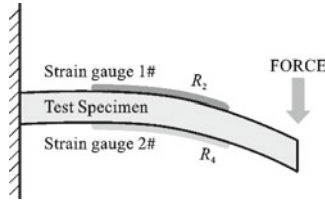
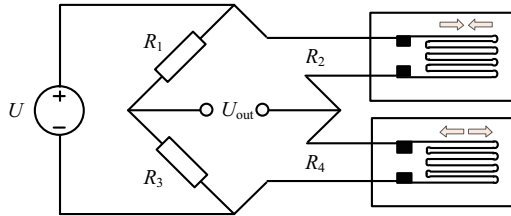
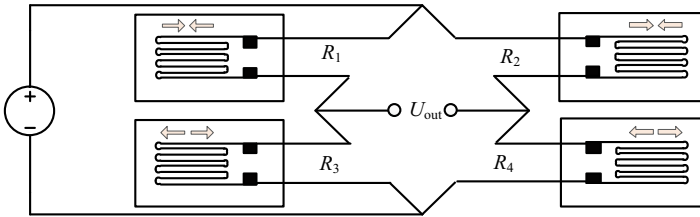


Fig. 7.15 Measuring strain with two strain gauges



(a) Half bridge



(b) full bridge

Fig. 7.16 Half bridge and full bridge for strain gauges

increased by replacing all the resistors in the bridge with strain gauges as shown in Fig. 7.16b.

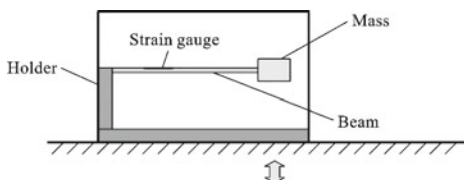
4. Applications of strain gauge

Usually, strain gauges are bonded to a beam-like structure with glue to sense its strain. With different conversion structures, many physical quantities, such as displacement and acceleration, can be converted into the strain in the beam. Thus, strain gauges have a lot of applications.

Example 7.5: Vibration Detector

Vibration detector can detect vibrations of an object or the ground. It has been used as a monitoring device to detect the destructive actions, such as digging walls, drilling holes and blasting in warehouses, banks and ancient buildings. Its structure is shown in Fig. 7.17. When the ground is vibrating, the holder will vibrate with it. Due to the

Fig. 7.17 Strain gauge vibration detector



inertia of the mass, the beam will experience strain. Connecting the strain gauge in a measurement circuit, then the vibration can be detected.

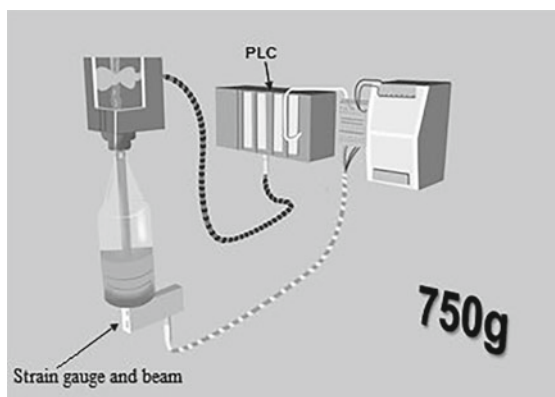
Example 7.6: Automatic Loading Sensor

In a modern beverage factory, the bottles are filled automatically. A sensor is needed to stop the filling of liquid when the bottle is full. Strain gauges can be used together with a beam to realize automatic loading as illustrated in Fig. 7.18. When filling the bottle, the weight of the bottle keeps increasing and bends the beam. The strain in the beam can be measured by a strain gauge, and can be used to infer the weight of the bottle. When the weight reaches the value set by the factory, the PLC (programmable logical controller) can send a signal to stop the filling.

DIY Experiment 7.2 Digital Scale

To build a digital scale, a beam with bonded strain gauge is needed. As we can see from Fig. 7.19, if we put a weight on the scale, it will bend the beam, and the strain caused by bending will be sensed by the strain gauge. The output voltage of the strain gauge is amplified and acquired by the Arduino board. Before using it to measure a weight, we need to calibrate the system by putting some standard weights on the scale and find the relationship between the weight and the output voltage.

Fig. 7.18 Automatic loading device based on strain gauge



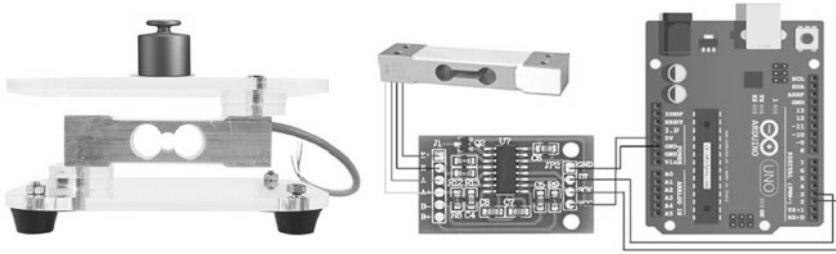


Fig. 7.19 Digital scale based on Arduino

7.2.3 Other Resistive Sensors

1. Resistance Thermometer

Resistance thermometer is also called resistance temperature detector (RTD). It is based on the fact that the resistance of metals is dependent on the ambient temperature. Their relationship can be modeled by the following equation:

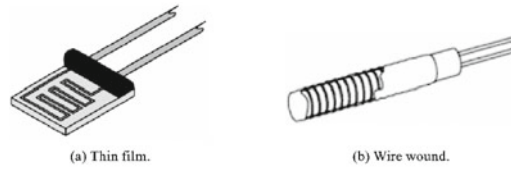
$$R_t = \begin{cases} R_0(1 + At + Bt^2) & (0^\circ\text{C} < t \leq 640^\circ\text{C}) \\ R_0[1 + At + Bt^2 + C(t - 100)t^3] & (-240^\circ\text{C} < t \leq 0^\circ\text{C}) \end{cases} \quad (7.20)$$

where t is the temperature in Celsius, R_t and R_0 are the resistances at $t^\circ\text{C}$ and 0°C , A , B and C are temperature coefficients, their values for copper are respectively $3.96847 \times 10^{-3}/^\circ\text{C}$, $-5.847 \times 10^{-7}/^\circ\text{C}$ and $-4.22 \times 10^{-12}/^\circ\text{C}$.

For most sensors made of metal, e.g. resistive strain gauge, we want the temperature confidants to be as small as possible to avoid the error caused by temperature variation during the measurement. However, if the sensor is particularly made to measure the temperature, the variation of temperature is reflected by the change of resistance. In this case, materials with large temperature coefficients, such as platinum, copper and nickel are preferably used to make the resistance thermometer. Among them, platinum is the most widely used material due to its high physical and chemical stabilities. Copper is easily oxidized when the temperature is higher than 100°C , so it is mainly used in the environment with relatively low temperature and without corrosive medium. The usage of nickel is limited by its difficulty of purification and bad linearity. There are also some new materials that are used in recent years for measuring ultra-low temperatures. For example, indium is used for temperatures between -269 to -258°C . Its sensitivity is 10 times higher than that of the platinum, but it has bad repeatability. Manganese is used for temperatures between -271 to -210°C . It also has a high sensitivity, but it is very brittle.

There are two main structures of the resistance thermometer as shown in Fig. 7.20. For the thin film type, a thin layer of the metallic material is deposited on an insulating substrate. And for the wire wound type, the metal wire is wound around an insulating core.

Fig. 7.20 Resistance thermometer structures



2. Thermistor

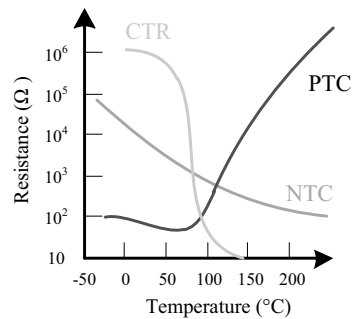
The term “thermistor” is the combination of “thermal” and “resistor”. Thermistor is also a sensor whose resistance changes with temperature. The difference between resistance thermometer and thermistor is that, resistance thermometer is made of metals while thermistor is made of metal-oxide, ceramic or polymer. According to the temperature characteristics shown in Fig. 7.21, thermistor can be divided into three categories:

- (1) Negative temperature coefficient (NTC) thermistor, whose resistance decrease as temperature increases.
- (2) Positive temperature coefficient (PTC) thermistor, whose resistance increases as temperature increases.
- (3) Critical temperature resistor (CTR), whose resistance drops suddenly at the critical temperature. Usually, it is used as a switch.

Comparing with resistance thermometer, thermistor has following advantages:

- (1) it has larger sensitivity, usually it is 10 times the sensitivity of a resistance thermometer;
- (2) it has a smaller size, so the temperature at a single point can be measured;
- (3) it has quicker response and it is more suitable for dynamic measurements;
- (4) it has higher stability, which can reach $0.0002\text{ }^{\circ}\text{C}$ within a measuring range of $0.01\text{ }^{\circ}\text{C}$;
- (5) it has lower power consumption.

Fig. 7.21 Temperature characteristics of thermistor



Thermistors have very wide usages in our daily life. For example, the working status of an air-conditioning is controlled by the temperature measured by thermistor. When the temperature reaches the value that we set, the air-conditioning will temporarily stop working to reduce the power consumption. Thermistor has also been used in refrigerators, rice cookers, electric ovens to measure the temperature and make further controls.

DIY Experiment 7.3 A Temperature Controlled Fan

In this experiment, we are going to design a temperature controlled fan. The schematic of the measuring and control system is shown in Fig. 7.22. The NTC thermistor and a resistor form a voltage divider. When the temperature increases, the analog voltage acquired by the Arduino board decreases. A relay is used to control the on and off of the motor that is connected to a fan. The relay module used in Fig. 7.22 is a low-level-trigger relay. When its input pin receives a low level voltage, the COM and NO contacts are closed. When the acquired voltage is less than a pre-set value (i.e. the temperature is higher than a specific value), the Arduino board will output a digital “LOW” to activate the relay and turn on the fan. A similar system has been used in laptops to reduce the CPU temperature when it is heating. A reference code is listed below. To use the code, the values used in the “if” statement should be adjusted according to the thermistor and resistor you choose.

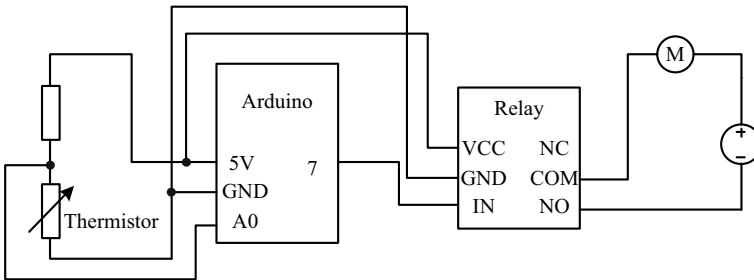


Fig. 7.22 Circuit for the temperature controlled fan

```
void setup() {  
    pinMode(7,OUTPUT);  
    pinMode(A0, INPUT);  
    digitalWrite(7,HIGH);  
}  
  
void loop() {  
    int val = analogRead(A0);  
    if (val<120){  
        digitalWrite(7,LOW);  
        delay(2000);  
    }  
    else {  
        digitalWrite(7,HIGH);  
        delay(2000);  
    }  
}
```

3. Photoresistor

Photoresistor is made of special semiconductors. When the material is exposed to light quantum, it absorbs energy and releases electrons, which increases the carrier density and mobility, and leads to the increase of the electrical conductivity. With stronger light intensity, the resistance of the material will be lower. The change of resistance with illumination is shown in Fig. 7.23.

Each material has a special sensitive frequency. Cadmium sulfide (CdS) and cadmium selenide (CdSe) are suitable for measuring the intensity of visible light, while zinc oxide (ZnO) and zinc sulfide (ZnS) are suitable for ultraviolet light, and lead sulfide (PbS), lead selenide (PbSe) and lead telluride (PbTe) are suitable for infrared light. A typical photoresistor is shown in Fig. 7.24, where the CdS material is coated onto the ceramic base in grids.

Example 7.7: Photoresistor-Based Counting System

Counting system is very important in the industry. For example, we can use it to count the total products that have been manufactured. A counting system can be designed as shown in Fig. 7.25a. Lasers are generated and directed towards the photoresistor.

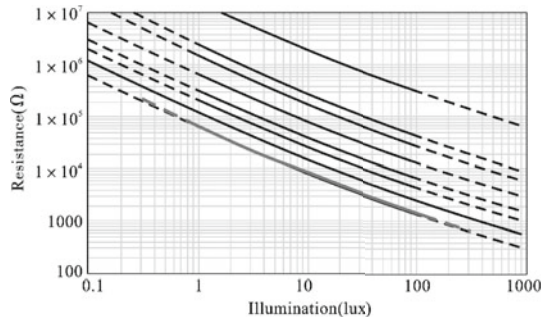


Fig. 7.23 Change of photoresistance with illumination

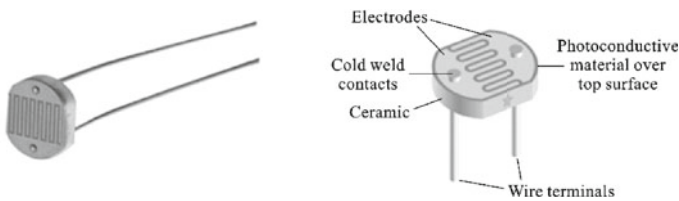


Fig. 7.24 A photoresistor with coated CdS

When an object passes by, it blocks some of the light and change the resistance of the photoresistor. The acquired signal would have the waveform shown in Fig. 7.25b. By counting the number of pulses with certain algorithm, we can easily know the number of pieces that have passed.

DIY Experiment 7.4 Control LED Lights with a Photoresistor

In this experiment, the number of LED lights that are turned on is controlled by the intensity of light in the room. A photoresistor is connected with a 10 k Ω resistor in series to form a voltage divider. The output voltage of the voltage divider is affected by the intensity of light and it is connected to an analog input pin of the Arduino board (Fig. 7.26). The acquired signal is compared with pre-set values, and if the intensity of light decreases, more LED lights are turned on. A reference code is listed below. To use this code, the output voltage of the voltage divider should be connected

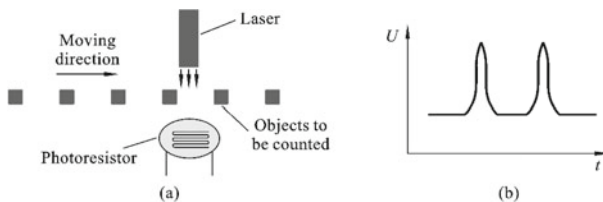
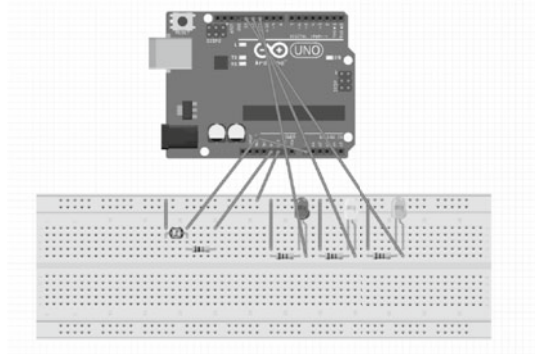


Fig. 7.25 A counting system with photoresistor

Fig. 7.26 Arduino board connection for the LED controlling system



to pin A0. Pins 11, 12 and 13 should be used to power the LED lights. The values used in the “if” statement should be adjusted according to the photoresistor and the brightness of the room.

```
int val = 0;

void setup() {
    pinMode(11,OUTPUT);
    pinMode(12,OUTPUT);
    pinMode(13,OUTPUT);
}

void loop() {
    val =  analogRead(A0);
    if (val <= 220){
        digitalWrite(11,LOW);
        digitalWrite(12,LOW);
        digitalWrite(13,LOW);
    }
    else if (val > 220 && val <= 350){
```



```

        digitalWrite(11,HIGH);
        digitalWrite(12,LOW);
        digitalWrite(13,LOW);
    }

    else if (val > 350 && val <= 500){
        digitalWrite(11,HIGH);
        digitalWrite(12,HIGH);
        digitalWrite(13,LOW);
    }

    else {
        digitalWrite(11,HIGH);
        digitalWrite(12,HIGH);
        digitalWrite(13,HIGH);
    }

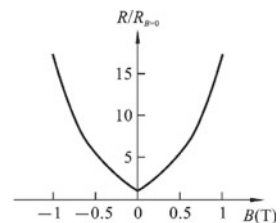
    delay(200);
}

```

4. Magnetoresistor

The resistance of certain semiconductors, such as InSb or InAs, changes with external magnetic field. The change of resistance with magnetic field for the InSb-NiSb magnetoresistor is shown in Fig. 7.27. This phenomenon is called magnetoresistive effect. The moving direction of carriers changes under the magnetic field, which reduces the current in the direction of external electric field. The reduction of current is equivalent to the increase of resistance. To have detectable magnetoresistive effect, the resistivity and carrier mobility of the material should be large enough.

Fig. 7.27 Characteristic curve of InSb-NiSb magnetoresistor



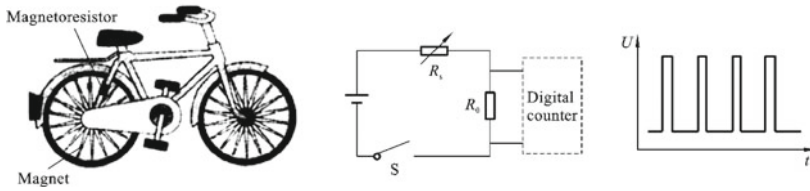


Fig. 7.28 Measuring bicycle speed by magnetoresistor

Example 7.8: Digital Compass

Nowadays, most smart phones have a digital compass for navigation. Magnetoresistors are used to measure magnetic field in all the directions. The direction with largest magnetic field is the direction of the earth magnetic field. The direction can be further read by the microcontroller and displayed on the screen.

Example 7.9: Measuring Bicycle Speed by Magnetoresistor

To measure the bicycle speed, a magnet is attached to the wheel and a magnetoresistor is fixed to the bicycle. The wheel rotates as the bicycle moves forward. Therefore, the magnet passes by the magnetoresistor periodically, and the magnetoresistor outputs periodic pulses as shown in Fig. 7.28.

5. Gas sensors

We are living in a world filled with gases. The composition and concentration of air seriously influence our health. For example, if the concentration of carbon monoxide reaches 0.8–1.15 mL/L, shortness of breath, rapid pulse, and even syncope may occur. Gas sensors are gas sensitive resistors, which can be used to measure the concentration of various gases.

Gas sensor is made of metal oxide such as SnO_2 , ZnO and MnO_2 . When absorbing flammable gases, such as hydrogen, carbon monoxide, alkanes, ethers, alcohols and natural gas, a chemical reaction will occur. It releases heat and increase the temperature. The resistance of the element will change due to the heating. There are many different types of gas sensors that are sensitivity to different gases as shown in Fig. 7.29. As can be seen from the figure, the MQ-5 sensor is more sensitive to CH_4 , thus it can be used to measure gas leakage in the kitchen. The MQ-3 sensor is more sensitive to alcohol, so it can be used for the test of drunk driving.

7.3 Inductive Sensors

Inductive sensors are based on the principle of electromagnetic induction. They are used to convert the measured quantity into the change of self-inductance of mutual-inductance. Inductive sensors are mainly used to measure displacement, and some other quantities that can be converted into displacement such as velocity, acceleration, pressure and weight.

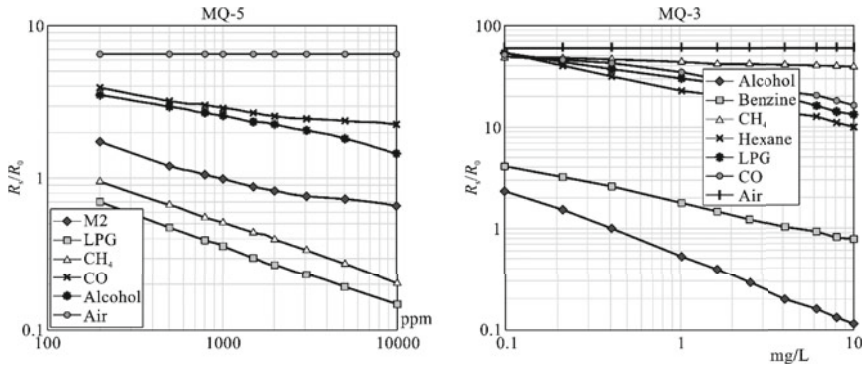


Fig. 7.29 Change of resistance with gas concentration

7.3.1 Self-inductive Sensor

1. Variable air gap type

The basic structure of the variable air gap type self-inductive sensor is depicted in Fig. 7.30. It consists of an armature and an iron core wound by a coil. The armature is connected to the structure to be measured and can move vertically. As the armature moves, the air gap between the armature and iron core changes, which further changes the reluctance of the magnetic circuit and the inductance of the coil.

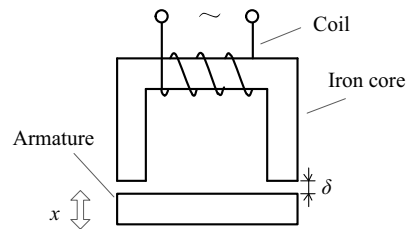
According to the Ohm's law for magnetic circuit, the magnetic flux is equal to the magnetomotive force divided by the reluctance:

$$\phi = \frac{E_m}{R_m} = \frac{NI}{R_m} \quad (7.21)$$

where ϕ is the magnetic flux, E_m is the magnetomotive force, R_m is the reluctance, N is the number of turns of the coil and I is the current in the coil. Combining with the definition of inductance, we get:

$$L = \frac{N\phi}{I} = \frac{N^2}{R_m} \quad (7.22)$$

Fig. 7.30 Structure of the variable air gap self-inductive sensor



where L is inductance of the coil.

In the magnetic circuit, the armature, iron core and air gap are connected in series, so the total reluctance is:

$$R_m = \frac{l_1}{\mu_1 S_1} + \frac{l_2}{\mu_2 S_2} + \frac{2\delta}{\mu_0 S} \quad (7.23)$$

where l_1 , l_2 and δ are the length of the iron core, armature and air gap, S_1 , S_2 and S are the cross-sectional area of the iron core, armature and air gap, μ_1 , μ_2 and μ_0 are the permeability of the iron core, armature and air. The iron core and armature are usually made of ferromagnetic materials with very high permeability (several hundreds or a thousand times of the vacuum permeability), thus their reluctances are small enough to be ignored. The reluctance in Eq. (7.23) can be approximated as:

$$R_m = \frac{2\delta}{\mu_0 S} \quad (7.24)$$

Substituting Eq. (7.24) into Eq. (7.22), the relationship between inductance and air gap length can be finally obtained:

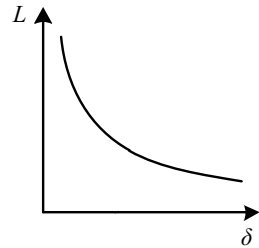
$$L = \frac{N^2 \mu_0 S}{2\delta} \quad (7.25)$$

From Eq. (7.25), we can find that the inductance is directly related to the length of the air gap. As the armature moves with the measured quantity, the air gap and corresponding inductance change accordingly. The change of inductance with air gap can be plotted according to Eq. (7.25) and it is shown in Fig. 7.31. It can be noticed from the figure that the inductance changes with air gap non-linearly, and this is a major drawback of the variable air gap type self-inductive sensor.

When the armature moves downwards and the air gap changes from δ_0 to $\delta_0 + \Delta\delta$, the corresponding change of inductance is:

$$\Delta L = \frac{N^2 \mu_0 S}{2(\delta_0 + \Delta\delta)} - \frac{N^2 \mu_0 S}{2\delta_0} = -L_0 \frac{\Delta\delta}{\delta_0} \frac{1}{1 + \Delta\delta/\delta_0} \quad (7.26)$$

Fig. 7.31 Change of inductance with air gap



where L_0 represents the original inductance when the air gap is δ_0 . Taylor series can be used to expand Eq. (7.26) to:

$$\Delta L = -L_0 \frac{\Delta \delta}{\delta_0} \left[1 - \frac{\Delta \delta}{\delta_0} + \left(\frac{\Delta \delta}{\delta_0} \right)^2 - \left(\frac{\Delta \delta}{\delta_0} \right)^3 + \dots \right] \quad (7.27)$$

From Eq. (7.27), we can also notice that the inductance changes non-linearly with air gap. In order to have a better linearity, the terms with higher orders must be ignored. To ignore those terms, the change of air gap should be much smaller than the original air gap

$$\Delta \delta \ll \delta_0 \quad (7.28)$$

Under this condition, Eq. (7.27) can be approximated with a linear relationship:

$$\Delta L \approx -\frac{L_0}{\delta_0} \Delta \delta \quad (7.29)$$

The sensitivity of the sensor can be further obtained:

$$K_\delta = \frac{\Delta L}{\Delta \delta} = -\frac{N^2 \mu_0 S}{2\delta_0^2} \quad (7.30)$$

Considering both Eq. (7.28) and Eq. (7.30), there is a trade-off when selecting the initial air gap δ_0 . To satisfy the condition in Eq. (7.28), the initial air gap should be large enough to ensure the linearity, while larger initial air gap will reduce the sensitivity as indicated in Eq. (7.30). The common values range from 0.1 to 0.5 mm.

2. Variable area type

Although the variable air gap type sensor can be used to measure displacement, it has the disadvantage of non-linearity. It can be noticed from Eq. (7.25) that the inductance is not only related to the length of air gap but also the cross-sectional area, and they have a linear relationship. The variable area type inductive sensor is shown in Fig. 7.32 with the armature moves horizontally to change the overlapped area of armature and iron core (i.e. the cross-sectional area of the magnetic flux). To get the change of inductance with area, Eq. (7.25) can be re-written in the following form:

$$L = \frac{N^2 \mu_0}{2\delta} S = K_S S \quad (7.31)$$

where K_S is the sensitivity.

3. Differential type

Fig. 7.32 Structure of the variable area type

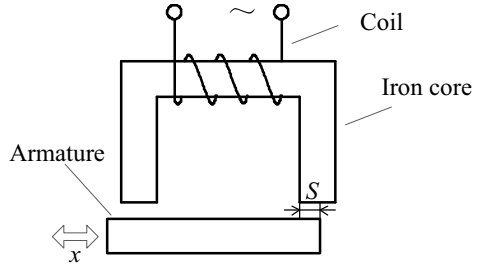
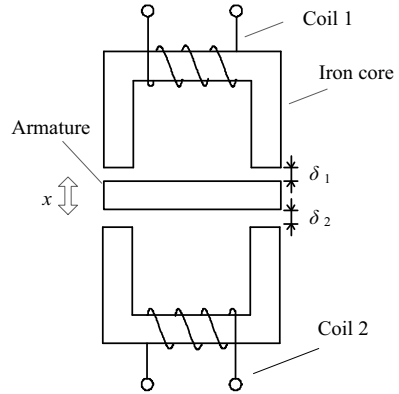


Fig. 7.33 Structure of the differential type



Another way to reduce the non-linearity in variable air gap sensor is to use a differential structure as shown in Fig. 7.33. The difference of the inductances of the two coils is used as the output. The two magnetic circuits are identical, thus the output inductance is zero at the initial balancing state.

When the armature moves upwards, the inductance of one coil increases while the other one decreases. The difference in inductance is:

$$\begin{aligned}
 \Delta L &= L_1 - L_2 \\
 &= \frac{N^2 \mu_0 S}{2(\delta_0 - \Delta \delta)} - \frac{N^2 \mu_0 S}{2(\delta_0 + \Delta \delta)} \\
 &= 2L_0 \frac{\Delta \delta}{\delta_0} \left[1 + \left(\frac{\Delta \delta}{\delta_0} \right)^2 + \left(\frac{\Delta \delta}{\delta_0} \right)^4 + \left(\frac{\Delta \delta}{\delta_0} \right)^6 + \dots \right] \\
 &\approx 2L_0 \frac{\Delta \delta}{\delta_0}
 \end{aligned} \tag{7.32}$$

where L_1, L_2 are the inductance of Coil 1 and Coil 2, δ_0 is the initial air gap length, $\Delta \delta$ is the displacement of the armature, L_0 is the inductance of one coil at the initial state. The terms in the Taylor series in Eq. (7.32) have higher orders than the terms in Eq. (7.27), so they can be ignored more easily. Therefore, the differential type

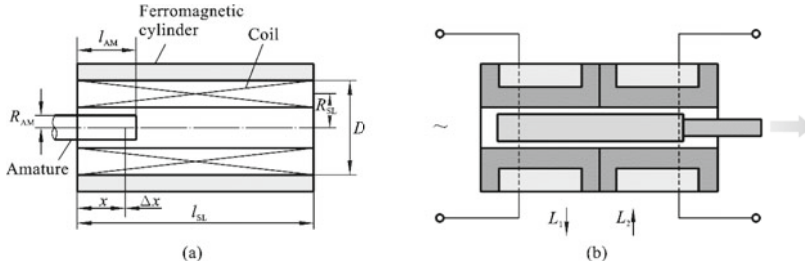


Fig. 7.34 Structure of the solenoid type self-inductive sensor

has better linearity than the variable air gap type. Besides, the sensitivity of the differential type is also doubled. The differential structure also helps to resist the interferences caused by temperature drift and voltage fluctuation of power supply.

4. Solenoid type

The structure of the solenoid type self-inductive sensor is shown in Fig. 7.34a. As the armature moves inside the solenoid, the inductance of the coil will change. Its inductance can be calculated by the Ohm's law for magnetic circuit, and the result is:

$$L = N^2 \left[\frac{\pi R_{SL}^2 \mu_0}{l_{SL} - l_{AM}} + \frac{2\pi \mu_0}{\ln(R_{SL}/R_{AM})} \frac{l_{AM}^3}{3l_{SL}^2} \right] \quad (7.33)$$

where L is the inductance of the solenoid, R_{SL} , l_{SL} are the radius and length of the solenoid, R_{AM} is the radius of the armature, l_{AM} length of the armature part in the coil. A differential structure can also be used for the solenoid type self-inductive sensor as shown in Fig. 7.34b.

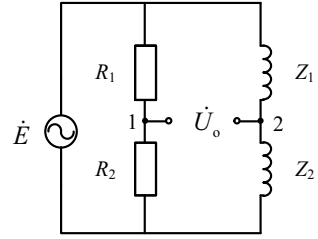
5. Measurement circuits for self-inductive sensors

Previous analyses of various self-inductive sensors relate the inductance with the displacement. To observe the change of inductance, measurement circuits are needed to convert the change of inductance into the change of voltage. The most commonly used circuit is the AC Wheatstone bridge shown in Fig. 7.35, in which Z_1 and Z_2 are the impedance of the two coils in differential type sensor, R_1 and R_2 are two resistors with the same resistance $R_1 = R_2 = R$. When the armature is in the initial balancing position, the impedances of two coils are the same, thus the output voltage is zero.

When the armature moves upwards, the inductance of Coil 1 increases by ΔL , while the inductance of Coil 2 decreases by ΔL . Then the output voltage is:

$$\begin{aligned} \dot{U}_o &= \dot{U}_1 - \dot{U}_2 \\ &= \frac{\dot{E}}{2} - \dot{E} \frac{Z_2}{Z_1 + Z_2} \end{aligned}$$

Fig. 7.35 AC Wheatstone bridge



$$\dot{U}_o = \frac{\dot{E}}{2} \frac{\Delta Z}{Z} \quad (7.34)$$

where \dot{U}_o is the output voltage, \dot{U}_1 and \dot{U}_2 are voltages at nodes 1 and 2, \dot{E} is the AC excitation, Z is the impedance of coils when the armature is at the initial position, which is expressed as:

$$Z = r_0 + j\omega L_0$$

where r_0 and L_0 are the resistance of inductance for the coils when the armature is at the initial position. For the self-inductive sensors, the change of impedance is mainly caused by the change of inductance, then Eq. (7.34) can be written as:

$$\dot{U}_o = \frac{\dot{E}}{2} \frac{j\omega \Delta L}{r_0 + j\omega L_0} = \frac{\dot{E}}{2} \frac{\omega \Delta L}{r_0^2 + \omega^2 L_0^2} (\omega L_0 + jr_0) \quad (7.35)$$

It can be noticed from Eq. (7.35) that the output voltage is a complex number. The imaginary part causes a phase shift between the output voltage and the excitation, which makes the analysis more difficult. Therefore, when designing the sensors, we should ensure the quality factor of the coil to be much larger than 1:

$$Q = \frac{\omega L_0}{r_0} \gg 1 \quad (7.36)$$

Under this condition, the imaginary part in Eq. (7.35) can be ignored. By further combining with Eq. (7.29), the relationship between the output voltage and displacement can be obtained:

$$\dot{U}_o = -\frac{1}{2\delta_0} \Delta \delta \dot{E} \quad (7.37)$$

Example 7.10: Proximity Switch Using Self-inductive Sensor

Proximity switch is a device that detects the proximity (closeness, or presence) of an object, and it is often used for the automatic control in industry. For example, in an automatic sorting system, proximity switch is used to sense the arrival of a workpiece,

then a mechanical arm is used to pick it. For the inductive proximity switch, the sensor only consists a coil and an iron core, and the workpiece is acting as the armature. When the workpiece is approaching the proximity switch, the inductance of the coil will change, and the change of inductance can be detected by a measuring circuit. Because the workpiece functions as the armature, the workpiece to be detected must be a ferromagnetic material such as steel.

Example 7.11: Measuring Rotational Speed of a Tooth Wheel

To monitor the operation status of a machine, the rotational speed of tooth wheel needs to be measured. A self-inductive sensor can be used to fulfill this task as shown in Fig. 7.36. Each tooth is acting as an armature in this example. As each tooth passes by, the distance between the armature and the iron core changes, which induces a periodic pulse signal. The rotational speed of the tooth wheel can be calculated as:

$$n_r = \frac{1}{T_{\text{pulse}} N_{\text{tooth}}}$$

where n_r is the rotational speed, T_{pulse} is the period of the pulse signal, N_{tooth} is the number of tooth in the tooth wheel.

DIY Experiment 7.5 Measuring the Gravitational Acceleration with Inductive Proximity Switch

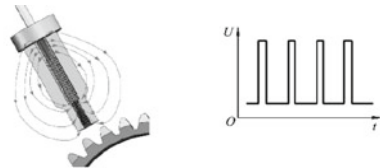
The period of a pendulum is related to the gravitational acceleration g and the length of the string that is attached to the mass L_s :

$$T_{\text{pend}} = 2\pi \sqrt{\frac{L_s}{g}}$$

Given a known string length, we only need to measure the period of a pendulum to calculate the gravitational acceleration. The experimental setup in Fig. 7.37 can be used to measure the period. As the ferromagnetic ball approaches the inductive proximity switch periodically, a periodic output voltage can be acquired by the Arduino board, and the gravitational acceleration can be calculated:

$$g = \frac{4\pi^2}{T_{\text{pend}}^2} L_s$$

Fig. 7.36 Measuring rotational speed of a tooth wheel using self-inductive sensor



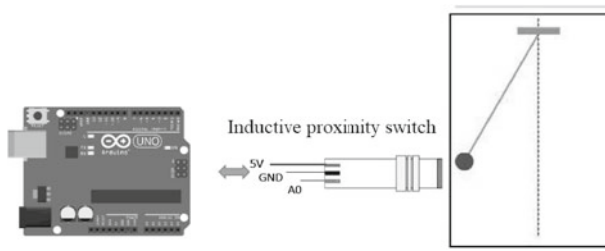


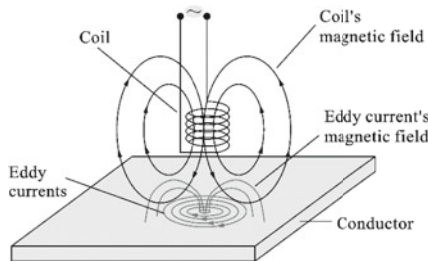
Fig. 7.37 Experimental setup for measuring the gravitational acceleration

7.3.2 Eddy Current Sensor

According to Faraday’s law of induction, when a conductor is placed in a space with time-varying magnetic field, an electromotive force and corresponding currents are generated inside the conductor. Because the currents flow in a closed vortex shape, it is called eddy current, and this phenomenon is called eddy current effect. The eddy current sensor is a coil, and its working principle is illustrated in Fig. 7.38. The coil carrying alternating current generates a changing magnetic field to induce eddy current in the conductive specimen, the eddy currents also generate a magnetic field that is opposing the primary magnetic field and changes the magnetic flux through the coil. As a result, the impedance of the coil will change due to the presence of eddy currents.

The intensity of eddy current is affected by many factors such as the distance between the coil and the specimen, the conductivity and magnetic permeability of the specimen and the frequency of the excitation signal. Eddy current sensors can be used to fulfill different tasks by changing one of the factors. For example, the eddy current sensor can be used as a displacement sensor similar to the self-inductive sensor. However, due to the different working principles, the specimen should be ferromagnetic material for the self-inductive sensor and conductive material for eddy current sensor respectively. Another typical application for the eddy current sensor is flaw detection. The flaws, such as crack and corrosion, inside a material change its conductivity and magnetic permeability and further change the impedance of

Fig. 7.38 Principle of eddy current sensor



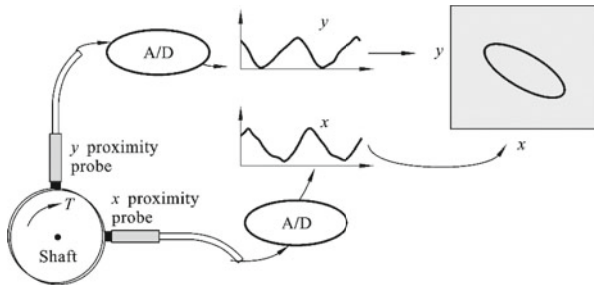


Fig. 7.39 Measuring shaft centerline orbit with eddy current sensors

the eddy current sensor. With a measuring circuit, we can observe the change of impedance by the output voltage.

Example 7.12: Measuring Shaft Centerline Orbit

Shaft is an essential component for many types of machinery. Fault in the shaft, such as the misalignment, could cause serious damage to the machine and bring economic losses. The working condition of a rotating shaft can be monitored with a pair of eddy current sensors as shown in Fig. 7.39. The two sensors are measuring the horizontal and vertical vibrations of the shaft respectively. The displacements in horizontal and vertical directions can be used as the abscissa and ordinate to plot the shaft orbit. By observing the shaft orbit, the type of fault can be diagnosed.

7.3.3 Mutual-Inductive Sensor

When two coils are brought in proximity, the change of current in one coil will induce electromotive force in the other coil:

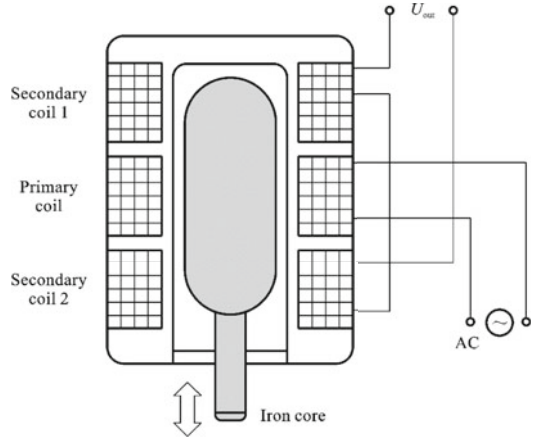
$$e_2 = -M \frac{di_1(t)}{dt} \quad (7.38)$$

where e_2 is the induced electromotive force in Coil 2, M is the mutual inductance and i_1 is the current in Coil 1. This phenomenon is called mutual induction and the mutual inductance is related to the relative position of the two coils and the magnetic permeability of the surrounding medium.

A typical mutual-inductive sensor is depicted in Fig. 7.40. The coil in the middle is connected to an AC power supply. The difference between the induced voltages in the other two coils is used as the output. The mutual-inductive sensor has a differential output and its working principle is similar to the transformer, hence it is also called differential transformer.

The equivalent circuit for the double solenoid type mutual-inductive sensor is shown in Fig. 7.41. The current flowing in the primary coil is:

Fig. 7.40 Double solenoid type mutual-inductive sensor



$$\dot{i}_1 = \frac{\dot{e}_1}{R_1 + j\omega L_1} \quad (7.39)$$

where \dot{e}_1 is the excitation voltage, R_1 and L_1 are the resistance and inductance of the primary coil respectively. Then the output voltage is:

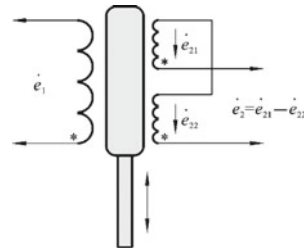
$$\dot{e}_2 = \dot{e}_{21} - \dot{e}_{22} = -j\omega(M_1 - M_2) \frac{\dot{e}_1}{R_1 + j\omega L_1} \quad (7.40)$$

where \dot{e}_2 is the output voltage, \dot{e}_{21} and \dot{e}_{22} are the voltages of the two secondary coils, M_1 and M_2 are the mutual inductance between the primary coil and the secondary coils. The amplitude of the output voltage can be further obtained:

$$e_2 = |\dot{e}_2| = \frac{\omega|M_1 - M_2|}{\sqrt{R_1^2 + \omega^2 L_1^2}} e_1 \quad (7.41)$$

When the iron core is in the initial position, the two secondary coils are symmetrical to the iron core, thus $M_1 = M_2$, and the output voltage is zero. As the iron core moves, one mutual inductance increases while the other decreases. The change of

Fig. 7.41 Equivalent circuit for double solenoid type mutual-inductive sensor



output voltage amplitude with displacement of iron core is plotted in Fig. 7.42. The amplitude is independent of the moving direction of the iron core. It can be noticed from Eq. (7.40) that the moving direction changes the phase of the output voltage by 180° . Thus, to find the moving direction, a phase detector is needed.

Example 7.13: Measuring Roughness with Differential Transformer

Measuring surface roughness is important for guarantying the precision of manufactured products. The structure depicted in Fig. 7.43 can be used to measure roughness by measuring displacement. The iron core of the differential transformer is connected to a pin by a support rod. A spring is used to make the iron core and pin go back to their original positions after the pin is pushed inside by the surface roughness. As the sensor moves above a rough surface, the iron core moves with respect to the surface, and the output voltage of the sensor can be used to indicate the surface roughness.

Example 7.14: Measure Pressure with Differential Transformer

Pressure measurement is important for some gas storage tanks and transmission pipelines to ensure there is no leakage. In the cars, there are also pressure gauges to monitor the tire pressure. The differential transformer can be used to measure the pressure as shown in Fig. 7.44. The membrane moves according to the pressure difference between two chambers. Since the membrane is connected to the iron core of a differential transformer by a linking rod, its movement will change the output of the differential transformer.

Fig. 7.42 Output voltage amplitude of the double solenoid type mutual-inductive sensor

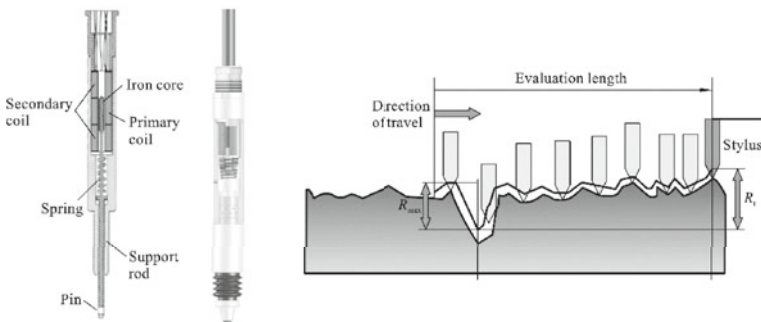
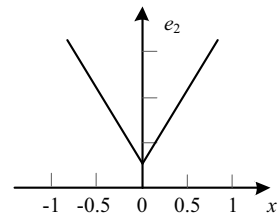
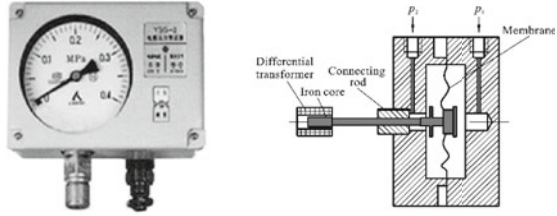


Fig. 7.43 Measuring roughness with differential transformer

Fig. 7.44 Pressure gauge using differential transformer



7.4 Capacitive Sensors

Capacitive sensor is a device that converts the measured physical quantity into the change of capacitance. A typical parallel-plate capacitor consists of two metal plates insulated by a dielectric medium is depicted in Fig. 7.45. The capacitance of the capacitor is:

$$C = \frac{\varepsilon_0 \varepsilon_r S}{\delta} \quad (7.42)$$

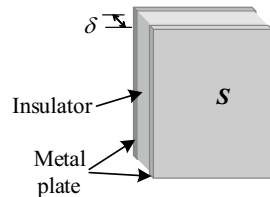
where C is the capacitance, ε_0 and ε_r are the vacuum permittivity and relative permittivity of the insulating medium, δ is the distance between the metal plates, S is the area of the metal plates. It indicates that any change in ε_r , δ , and S would cause a change in capacitance. We can use capacitive sensor to measure different physical quantities by keeping two of the parameters and change the other one.

According to Eq. (7.42), the capacitance changes with the distance between the metal plates. The capacitance and distance have a nonlinear relationship and the sensitivity changes with distance:

$$K_\delta = \frac{dC}{d\delta} = -\varepsilon_0 \varepsilon_r S \frac{1}{\delta^2} \quad (7.43)$$

It can be noticed from Eq. (7.43) that the sensitivity decreases with distance. Thus, similar to the variable air gap inductive sensor, the variable distance capacitive sensor is used to measure small distances. Because of the non-linearity, the change of distance should be small, usually is $\Delta\delta/\delta_0 \approx 0.1$. In practical application, a differential structure is used to increase the linearity and resist temperature drift and voltage fluctuation of the power supply as illustrated in Fig. 7.46.

Fig. 7.45 Configuration of a capacitor



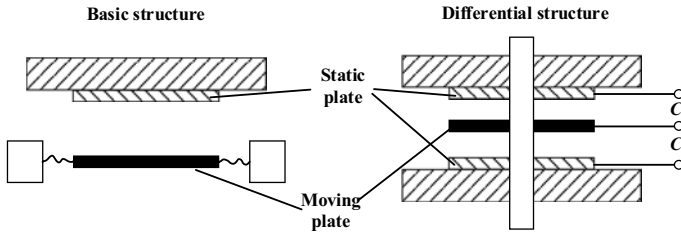


Fig. 7.46 Comparison of the basic and differential structures of the variable distance type capacitive sensor

The capacitance is also related to the area of metal plates. By changing the overlapped area of the two plates as shown in Fig. 7.47, we can make the variable area type capacitive sensors. Figure 7.47a shows a linear displacement sensor, which has a fixed plate and a movable plate. As the moving plate moves in the x -axis, the overlapped area and corresponding capacitance changes:

$$C = \frac{\varepsilon_0 \varepsilon_r b}{\delta} x \quad (7.44)$$

where x is the length of the overlapped area, b is the width of the plate. Figure 7.47b shows a linear displacement sensor with coaxial cylindrical metals. Its capacitance is:

$$C = \frac{2\pi \varepsilon_0 \varepsilon_r}{\ln(D/d)} x \quad (7.45)$$

where D and d are the diameters of the external and internal cylinders respectively. Figure 7.47c shows an angular displacement sensor, its capacitance changes with the overlapped angle:

$$C = \frac{\varepsilon_0 \varepsilon_r r^2}{2\delta} \theta \quad (7.46)$$

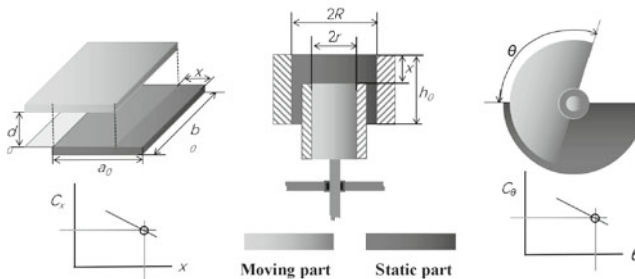


Fig. 7.47 Variable area type capacitive sensors

where θ is the overlapped angle.

As we can see from Eqs. (7.44) to (7.46), the advantage of the variable area type capacitive sensor is the linear relationship between input and output. However, it has lower sensitivity than the variable distance type capacitive sensor. It is suitable for measuring large displacement.

The capacitance is also sensitive to the permittivity (also called dielectric constant) of the insulating medium. Thus, we can use the capacitive sensor to measure the change of material property between the two plates. The relative permittivity of some selected materials is listed in Table 7.3. Based on this effect, it can be used to measure physical quantities such as humidity, temperature and liquid level.

Example 7.15: Liquid Level Gauge with Capacitive Sensor

In Example 7.3, a resistive sensor is used to detect the gas reservation by measuring the liquid level. In this example, we will show a different way to measure liquid level using a capacitive sensor. Its principle is depicted in Fig. 7.48. When the liquid level rises, the air between the two metal plates is gradually replacing by water, which changes the permittivity of insulating medium and the capacitance of the capacitor.

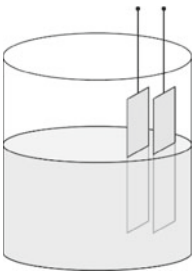
Example 7.16: Capacitive Touchscreen

As we have already mentioned in Example 7.4, capacitive touchscreen is replacing resistive touchscreen in smart phones and many home appliances. The principle of capacitive touch sensor is illustrated in Fig. 7.49. As our finger touches the screen,

Table 7.3 Relative permittivity of common materials

Material		Relative permittivity
Air		1.0006
Water	0 °C	87.9
	20 °C	80.2
	100 °C	55.5
Rubber		7
Teflon		2.1
Concrete		4.5
Glass		3.7–10

Fig. 7.48 Liquid level gauge based on capacitive sensor



it acts as another conductive plate and introduces a capacitance that is parallel to the original sensor capacitance and change the total capacitance.

Example 7.17: Capacitive Proximity Switch

From previous analyses we know that the capacitance is dependent on the permittivity of the medium between two electrodes. A capacitive proximity switch can be made to detect the presence of an object based on this effect, and its structure is shown in Fig. 7.50. To make it easier to detect objects nearby, the electrodes of the capacitor are placed in the same plane. The capacitor is further connected to an inductor to form an L - C oscillator. The frequency of the output signal changes with the capacitance, and is measured by a frequency detector. The capacitive proximity switch can be used for both non-metallic materials (e.g. water, oil, paper, wood, glass, etc.) and metals (e.g. copper, aluminum, steel, etc.). When a non-metallic material is nearby, it changes the permittivity of the medium between electrodes, hence changes the capacitance. When a metal is nearby, it is acting as another electrode and introduce a parallel capacitor to change the total capacitance.

Example 7.18: Condenser Microphone

Condenser is an obsolete term for capacitor. But, when describing the capacitor-based microphone, “condenser microphone” is more commonly used. Its structure is depicted in Fig. 7.51. A variable distance type capacitor is used, which has a fixed back plate and a diaphragm that moves under the struck of sound waves. The capacitor is connected to a DC voltage source and a resistor in series. The electrical charge held on the capacitor is:

Fig. 7.49 Principle of capacitive touch sensor

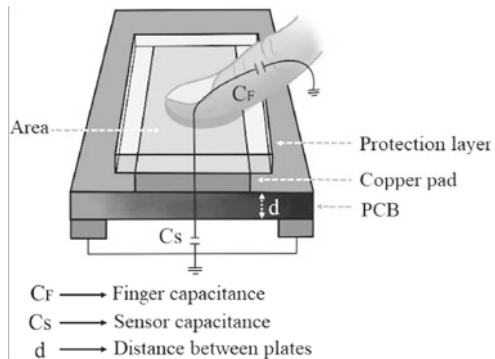


Fig. 7.50 Structure of capacitive proximity switch

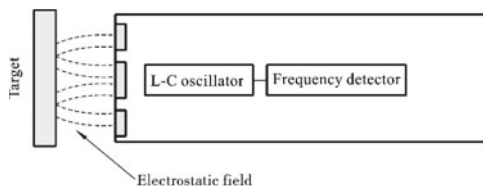
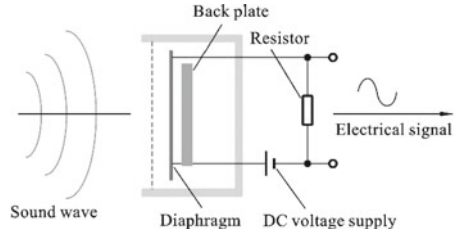


Fig. 7.51 Structure of a condenser microphone



$$Q_C = CV \quad (7.47)$$

where C is the capacitance and V is the electrical potential across the capacitor. When the diaphragm moves closer to the back plate under sound pressure, the capacitance increases and the capacitor is being charged. During the charging process, current flows across the resistor to generate an electrical signal. When the diaphragm moves away from the back plate, the discharging of the capacitor also induces a current in the circuit.

DIY Experiment 7.6 Simulate the Automatic Windscreen Wiper of a Car

Windscreen wiper is an indispensable device in cars to remove rain, ice and dust from the front window of a car. Nowadays, most cars have automatic windscreen wiper that is turned on automatically when detecting water on the windscreen. We can try to simulate an automatic windscreen wiper using a capacitive humidity sensor, a servo, plastic sticks and an Arduino board. When there is water on the humidity sensor, the change in permittivity causes the change of capacitance. The servo starts to rotate the plastic sticks after detecting the presence of water.

7.5 Magnetoelectric Transducer

Magnetoelectric transducer is a device that converts measured physical quantity into electromotive force. According to Faraday's law of induction, the magnitude of the induced electromotive force in any closed loop is equal to the rate of change of the magnetic flux through this loop:

$$e = -N \frac{d\phi}{dt} \quad (7.48)$$

where e is induced electromotive force, N is the number of turns of the coil, ϕ is the magnetic flux through the loop, t is the time.

7.5.1 Moving Coil Type

The relative movement between coil and magnet would cause the change of magnetic flux in the coil. The moving coil type is shown in Fig. 7.52, it can be used to measure both translational and angular speeds.

For the magnetoelectric transducer shown in Fig. 7.52a, there is a constant magnetic field in the air gap. As the coil moves vertically, a motional electromotive force is induced:

$$e = NBlv \tag{7.49}$$

where B is magnetic flux density in the air gap, l is the length for one turn of the coil, N is the number of turns of the coil inside the air gap, v is the relative velocity between the coil and magnetic field. As we can see from Eq. (7.49), the induced electromotive force is proportional to the velocity. Thus, we can use it to measure velocity or vibration.

Example 7.19: Moving Coil Microphone

The structure of a moving coil microphone is displayed in Fig. 7.53. A coil is attached to a flexible diaphragm. As the sound wave arrives and struck the diaphragm, the coil will vibrate with the diaphragm. The relative movement between coil and magnet will generate an electrical signal.

Fig. 7.52 Moving coil type magnetoelectric transducer

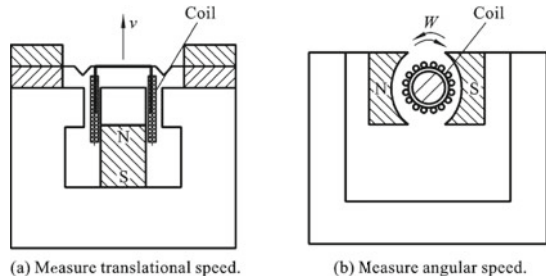


Fig. 7.53 Moving coil microphone

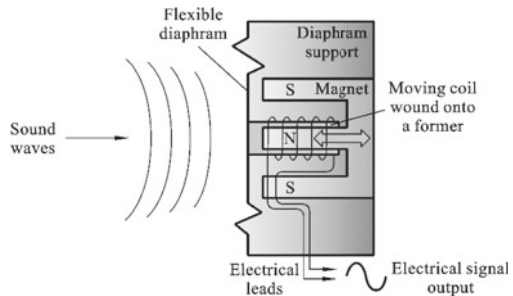
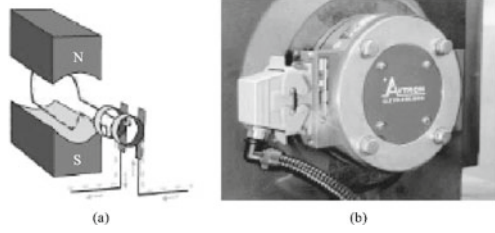


Fig. 7.54 A tachometer based on magnetolectric transducer



Example 7.20: Moving Coil Tachometer

Tachometer is an instrument which measures the rotational speed of a shaft or disc. The magnetolectric transducer displayed in Fig. 7.54a can be used to build a tachometer (Fig. 7.54b). The working principle of tachometer is similar to the generator. As the shaft is connected to the coil between two magnetic poles, an electromotive force is generated when the coil rotates with the shaft:

$$e = BS\omega \sin(\omega t) \quad (7.50)$$

where S is the area of the coil, ω is the rotational speed. The induced electromotive force is proportional to the rotational speed of the shaft.

7.5.2 Variable Reluctance Type

The magnetic flux through a coil is also dependent on the reluctance of magnetic circuit. Therefore, besides moving coil directly, the structures in Fig. 7.55 can be used to measure translational and rotational speeds by changing the reluctance. The reluctance of the magnetic circuit in Fig. 7.55a can be approximated by the reluctance of the air gap:

$$R_m = \frac{2\delta}{\mu_0 S} \quad (7.51)$$

where R_m is the reluctance of the magnetic circuit, δ is the air gap, S is the cross-sectional area. The magnetic flux can be calculated by the Ohm's law for magnetic circuit:

$$\phi = \frac{E_m}{R_m} = \frac{E_m \mu_0 S}{2\delta} \quad (7.52)$$

where E_m is the magnetomotive force. Then, the induced electromotive force is:

$$e = -N \frac{d\phi}{dt} = -N \frac{d\phi}{d\delta} \frac{d\delta}{dt} = \frac{N E_m \mu_0 S}{2\delta^2} v \quad (7.53)$$

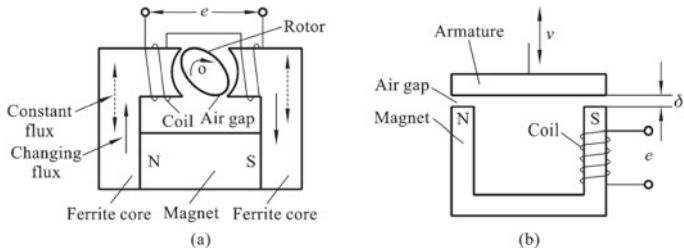


Fig. 7.55 Variable reluctance type magnetoelectric transducer

where v is the speed of the armature. It can be seen from Eq. (7.53) that the induced electromotive force is proportional to the armature speed.

At first glance, the variable reluctance type magnetoelectric transducer in Fig. 7.55a is quite similar to the variable air gap type inductive sensor in Fig. 7.30, whereas, there are some major differences between them. Firstly, the magnetoelectric transducer is an energy conversion type sensor, the energy of induced electromotive force comes directly from the kinetic energy of the armature. While the inductive sensor is an energy control type, the air gap distance only changes the inductance of the coil, and another power supply is needed to convert the change of inductance into the change of voltage. Secondly, the magnetoelectric transducer is only capable of measuring dynamic displacement. As can be seen from Eq. (7.53), the induced electromotive force is zero when the air gap δ is a constant. While the inductive sensor can be used for both static and dynamic measurement, its inductance is related to the air gap δ as indicated by Eq. (7.25). Therefore, the static displacement can be calculated by measuring the inductance with a circuit.

Example 7.21: Torque Measurement Using Magnetoelectric Transducer

An actuator is sometimes connected to a loading with a shaft as illustrated in Fig. 7.56. To measure the torque in the shaft, two magnetoelectric transducers can be used to monitor the rotation of the two tooth wheels on both sides of the shaft. The magnetoelectric transducer is consisted of a permanent magnet and a coil with ferrite core. As each tooth passes by, the reluctance is changed and an electromotive force is induced in the coil. When there is torque in the shaft, the signals acquired by the two magnetoelectric transducer will have a phase shift that is proportional to the torque. The phase shift can be detected with the circuit displayed in Fig. 7.57. The signals acquired by the magnetoelectric transducers are fed into the diode voltage limiter. Then, the signals are digitized by voltage comparators. Finally, an XOR gate is used to get their phase difference. The process of signal processing is shown in Fig. 7.57b.

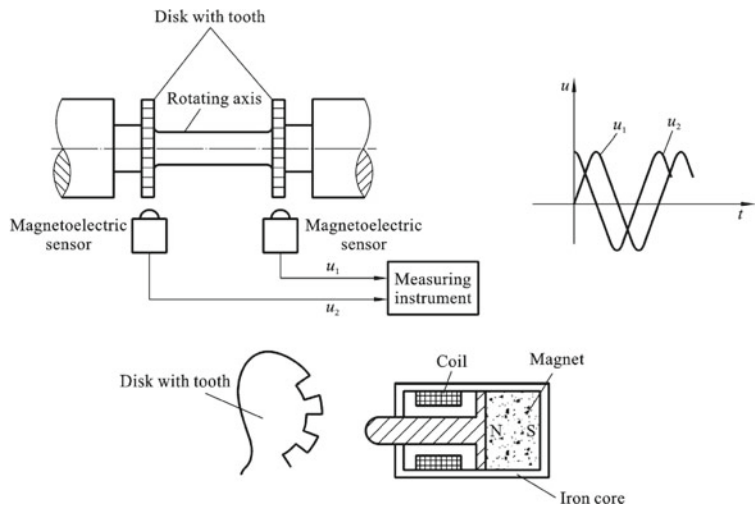


Fig. 7.56 Principle of torque measurement using magnetoelectric transducer

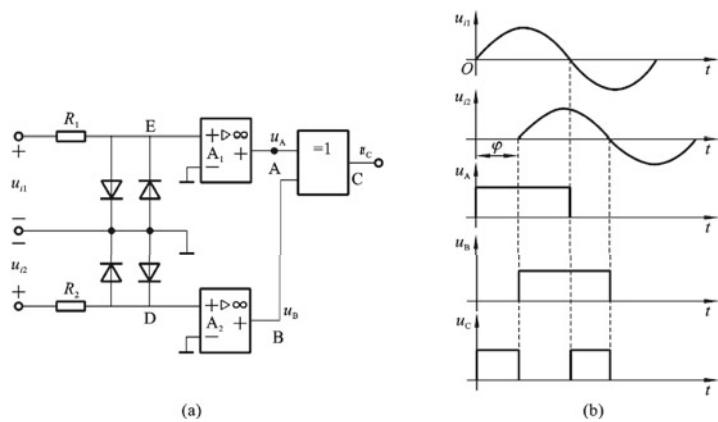


Fig. 7.57 Phase detection for the magnetoelectric transducer signals

7.6 Piezoelectric Transducer

7.6.1 Piezoelectric Element

Piezoelectric transducer is made based on the piezoelectric effect of certain solid materials (such as crystal). Piezoelectric effect refers to the accumulation of electric charges in certain solid materials in response to an externally applied mechanical stress. It was discovered by Jacques Curie and Pierre Curie. The electric charges will disappear once the external stress is removed. Conversely, when an electric field

is applied to the piezoelectric material, the dimensions of the material will change accordingly, and this is called converse piezoelectric effect.

The direct and converse piezoelectric effect can be mathematically expressed as Eqs. (7.54) and (7.54) respectively.

$$\mathbf{D} = \mathbf{d}\mathbf{T} + \varepsilon^T \mathbf{E} \quad (7.54)$$

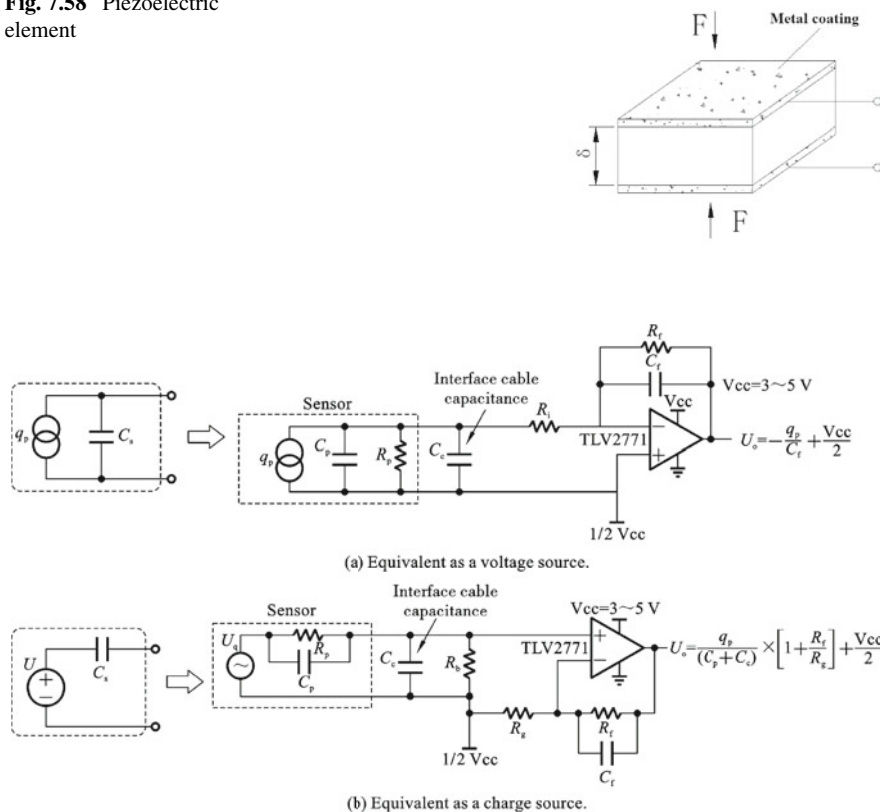
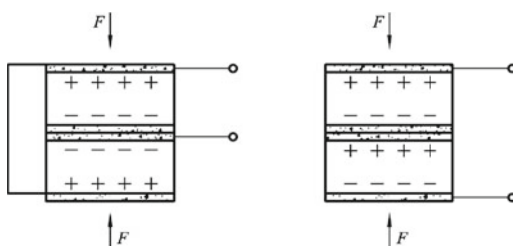
$$\mathbf{S} = \mathbf{s}^E \mathbf{T} + \mathbf{d}^T \mathbf{E} \quad (7.55)$$

where \mathbf{D} is the electric displacement field, \mathbf{d} is the matrix for direct piezoelectric effect, \mathbf{T} is the stress, ε^T is the permittivity, \mathbf{E} is the electric field strength, \mathbf{S} is the strain, \mathbf{s}^E is the compliance matrix, \mathbf{d}^T is the matrix for converse piezoelectric effect. Therefore, on one hand, the piezoelectric transducer can work as a sensor to convert the stress into electrical signal based on the direct piezoelectric effect. On the other hand, based on the converse piezoelectric effect, the piezoelectric transducer can also work as an actuator to convert the electrical signal to mechanical strain and generate acoustic waves. For a typical piezoelectric ceramic, the matrix elements in Eq. (7.55) can be explicitly expressed as:

$$\begin{bmatrix} S_{11} \\ S_{22} \\ S_{33} \\ S_{23} \\ S_{13} \\ S_{12} \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & s_{13} & 0 & 0 & 0 \\ s_{21} & s_{22} & s_{23} & 0 & 0 & 0 \\ s_{31} & s_{32} & s_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & s_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & s_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & s_{66} \end{bmatrix} \begin{bmatrix} T_{11} \\ T_{22} \\ T_{33} \\ T_{23} \\ T_{13} \\ T_{12} \end{bmatrix} + \begin{bmatrix} 0 & 0 & d_{31} \\ 0 & 0 & d_{32} \\ 0 & 0 & d_{33} \\ 0 & d_{24} & 0 \\ d_{15} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} E_{11} \\ E_{22} \\ E_{33} \end{bmatrix} \quad (7.56)$$

A piezoelectric element is a force-sensitive element with electrodes coated onto the piezoelectric material as depicted in Fig. 7.58. The piezoelectric element has the shape of a parallel plate capacitor, thus it has the capacitance of $C = \varepsilon_0 \varepsilon_r S / \delta$. When the piezoelectric element is acting as a sensor, the induced charge and voltage are inter-changeable by the equation $Q_C = CV$. Therefore, the piezoelectric element can be equivalent as a capacitor with a charge source or a capacitor with a voltage source in a circuit. Correspondingly, the charge amplifier and voltage amplifier can be used for the piezoelectric element as shown in Fig. 7.59.

Multiple piezoelectric elements can be used together by connecting each other in series or parallel as depicted in Fig. 7.60. When connected in parallel, the capacitance is larger, hence there are more electric charges induced. It is suitable for the situation where the piezoelectric element outputs electric charges. Besides, it is suitable for low frequency measurement since it has large time constant. When connected in series, the capacitance is small, so the induced electric charge is less and the induced voltage is larger. It is suitable for the situation where the piezoelectric element outputs voltage.

Fig. 7.58 Piezoelectric element**Fig. 7.59** Amplifier circuit for piezoelectric element**Fig. 7.60** Parallel and series connection of piezoelectric elements

The voltage signal of a piezoelectric transducer is generated by the electric charges. If the electric charge flows into the measurement circuit, the voltage will

disappear. Therefore, the measurement circuit for piezoelectric element should have high input impedance.

Example 7.22: Piezoelectric Accelerometer

A piezoelectric accelerometer is a device that converts acceleration into electric signals. Its structure is depicted in Fig. 7.61. Two piezoelectric elements are connected in parallel and fixed to the support. A spring is also attached to the support and presses a mass above the piezoelectric elements. As the support accelerates, a force of $F = ma$ is applied to the piezoelectric elements and it can be converted into electrical signals due to the piezoelectric effect.

Example 7.23: Smart Materials and Structural Health Monitoring

Smart materials are materials with embedded sensors which can monitor the integrity of the structure. It is especially desired in the aerospace industry. During flight, aircraft is struck by birds and small rocks between times. The impact would damage to aircraft structures and cause disasters such as plane crash. Researchers in aerospace industry are developing the smart materials now. The idea is to embed a network of piezoelectric elements in the composite materials during manufacturing. When an impact happens, guided waves are generated due to the impact force. The mechanical waves will propagate through the plate and be finally acquired by the embedded piezoelectric elements. A typical example of the guided wave signals is displayed in Fig. 7.62. By extracting the arrival time of guided wave in each sensor, the location of impact can be estimated. Then, nondestructive testing techniques can be applied to examine the region of impact to check if there is a defect. The piezoelectric elements can also fulfill the task of defect inspection by themselves. One of the piezoelectric elements needs to be used as actuator to generate another ultrasonic guided wave after the impact, and others will listen to the guided wave. If damage is present in the path of the guided wave, we can see a perturbation in the acquired signals.

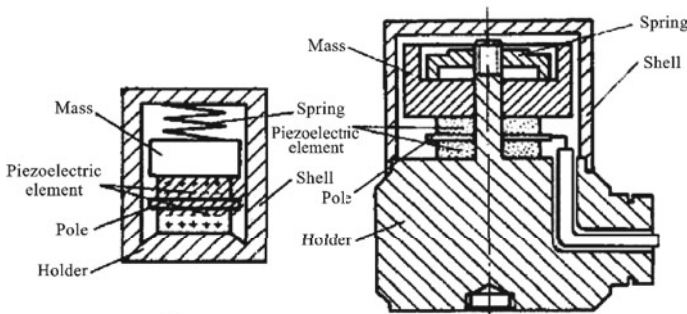


Fig. 7.61 Structure of a piezoelectric accelerometer

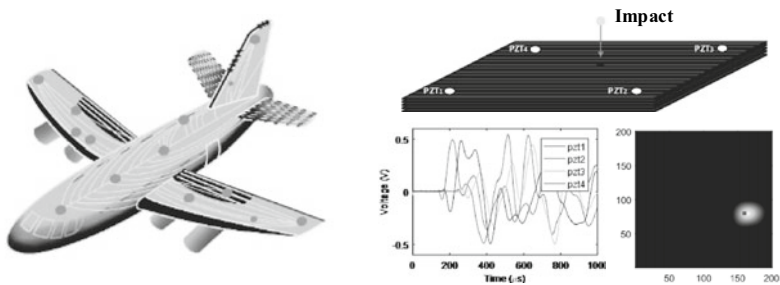


Fig. 7.62 Structural health monitoring using piezoelectric elements

7.6.2 Ultrasonic Transducer

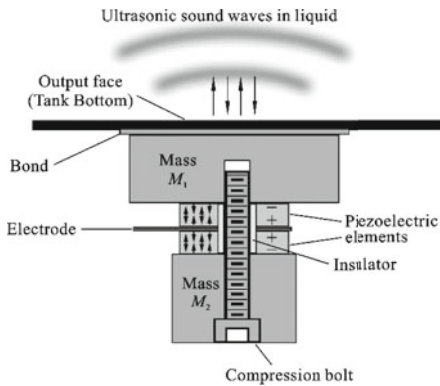
Ultrasonic transducer is a device that generates and receives ultrasound. There are typically three types of ultrasonic transducer according to the working principle: piezoelectric ultrasonic transducer, magnetostrictive ultrasonic transducer and electromagnetic ultrasonic transducer. The piezoelectric ultrasonic transducer is the most widely used ultrasonic transducer, and it is made of piezoelectric materials as shown in Fig. 7.63.

The ultrasound generated by the ultrasonic transducer is sound with a frequency higher than 20 kHz. It is defined with respect to the audible sound, which has a frequency between 20 Hz to 20 kHz. Since wavelength λ is related to frequency f by the equation:

$$\lambda = \frac{c}{f} \tag{7.57}$$

where c is the speed of sound, the ultrasound would have a smaller wavelength when compared with audible sound. Thus the ultrasound has better directivity and can transmit in narrow sound beams.

Fig. 7.63 Piezoelectric ultrasonic transducer



A main application of the ultrasonic transducer is thickness gauging, its principle is illustrated in Fig. 7.64. The ultrasonic transducer generates an ultrasonic wave that propagates in a straight line with constant speed until it encounters the interface of two media (i.e. the back surface). When ultrasonic wave meets the interface, a part of the energy will be reflected at the interface, while the other part is transmitted through the interface and continues propagating in the other medium. The reflection coefficient is dependent on the acoustic impedances of the two media:

$$R = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1} \right)^2 \quad (7.58)$$

where the acoustic impedance is defined as:

$$Z = \rho c \quad (7.59)$$

where ρ is the density and c is the sound speed. The reflected wave will be acquired by the ultrasonic transducer, and a typical signal is displayed in Fig. 7.62. By extracting the time-of-flight of the reflected wave Δt , the thickness can be calculated:

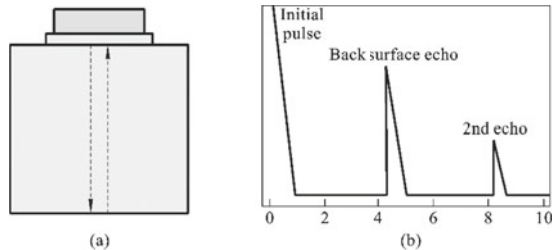
$$d = \Delta t \cdot c / 2 \quad (7.60)$$

The ultrasonic transducer can also be used for defect inspection based on a similar effect. When a defect exists in the material, it introduces another interface inside the material and causes a new echo as shown in Fig. 7.65. The depth of the defect can be determined by the time-of-flight as shown in Eq. (7.60) and the size can be estimated by the amplitude of the crack echo.

Example 7.24: Parking Radar

Nowadays, parking radars have been installed in most cars. It is capable of measuring the distance between the car and surrounding obstacles (walls, other cars, etc.), and makes an alarm when the car is going to hit an obstacle. The parking radar is based on a principle similar to ultrasonic thickness gauging. Instead, the ultrasonic wave travels in air and reflected at the obstacle. We can also use Eq. (7.60) to calculate the distance.

Fig. 7.64 Ultrasonic thickness gauging



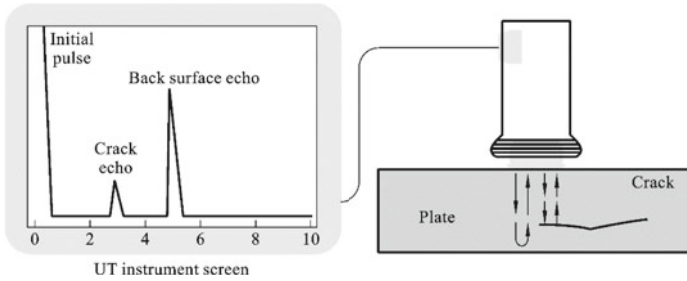


Fig. 7.65 Ultrasonic testing for defects

Example 7.25: Ultrasonic Flowmeter

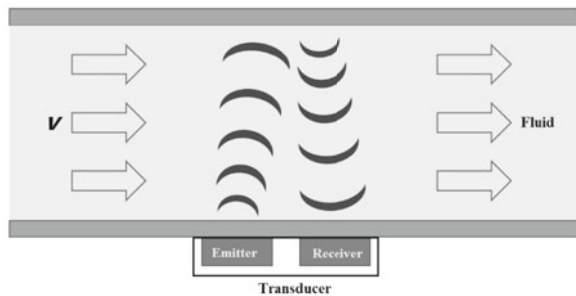
Flowmeter is a device which measures the flowing speed of fluid or gas. It is quite useful for detecting leakage of underground pipeline. After several years of burying underground, pipelines could be corroded. This may cause serious disasters such as oil leakage or even explosion. In an intact pipeline, oil flows at a constant speed in the whole pipeline. When leakage appears at some point, the flowing speed reduces after the leakage point. Thus, we can use a flowmeter to detect the leakage point. The principle of ultrasonic flowmeter is shown in Fig. 7.66. One ultrasonic transducer is working as an emitter to send the ultrasonic wave, and the other one receives the ultrasonic wave reflected from the pipe wall. According to the Doppler effect, the received ultrasonic wave has a frequency shift of

$$\Delta f = \frac{2f_0 v \cos(\theta)}{c - v \cos(\theta)} \quad (7.61)$$

where Δf is the frequency shift, f_0 is the original frequency of the ultrasonic wave, v is the speed of fluid, c is the speed of sound, θ is the angle between the directions of ultrasonic wave and fluid. When the speed of sound is much larger than the speed of fluid, Eq. (7.61) can be simplified as

$$\Delta f = 2f_0 v \cos(\theta)/c \quad (7.62)$$

Fig. 7.66 Ultrasonic flowmeter



From Eq. (7.62), we can find that the flowing speed of fluid is proportional to the frequency shift. By measuring the frequency of ultrasonic wave, the flowing speed can be determined.

DIY Experiment 7.7: Ultrasonic Ranging Device Based on Arduino

The principle of ultrasonic ranging has already been explained. With an ultrasonic ranging module (HC-SR04), we can build our own ultrasonic ranging device. The module has four pins: VCC, GND, TRIG, ECHO. The TRIG pin is connected to pin 8 of the Arduino board. When we give a low level voltage of $2\text{ }\mu\text{s}$ and high level voltage of $10\text{ }\mu\text{s}$, it will send an ultrasonic wave. The ECHO pin is connected to pin 7 of Arduino and we can use the pulseIn function to read the time-of-flight of the reflected ultrasonic wave. Then we can calculate the distance according to Eq. (7.60). Pin 9 of the Arduino is connected to an LED, which makes an alarm when the distance between the ultrasonic module and measured object is less than 5 cm. The reference code below needs a computer to display the measured distance. We recommend you try to add an LCD screen to display the distance and use a battery to power your board, so you can build a portable device as shown in Fig. 7.67.

```
float dis;
```

```
void setup() {  
    Serial.begin(9600);  
    pinMode(8,OUTPUT);  
    pinMode(7,INPUT);  
    pinMode(9,OUTPUT);  
}
```

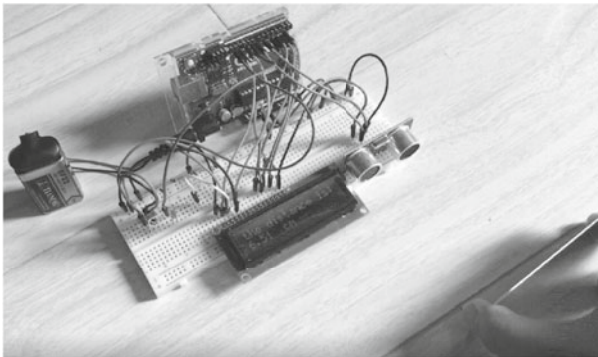


Fig. 7.67 Portable ultrasonic ranging device based on Arduino

```
void loop() {  
    digitalWrite(8,LOW);  
    delayMicroseconds(2);  
    digitalWrite(8,HIGH);  
    delayMicroseconds(10);  
    digitalWrite(8,LOW);  
  
    dis = pulseIn(7, HIGH) / 58.0;  
    dis = (int(dis*100.0))/100.0;  
    Serial.print("Distance = ");  
    Serial.print(dis);  
    Serial.println("cm");  
    if(dis < 5){  
        digitalWrite(9,HIGH);  
        delay(200);  
    }  
    else{  
        digitalWrite(9,LOW);  
        delay(200);  
    }  
    delay(500);  
}
```

7.6.3 QCM Humidity Sensor

QCM (Quartz Crystal Microbalance) is a quality inspection instrument that converts the change of surface quality of the electrode into the change of frequency, and can achieve detection in the magnitude of nanogram. It is currently the most widely used piezoelectric resonance sensor. QCM follows the inverse piezoelectric effect. When an alternating electric field is applied to both ends of the QCM, the QCM will produce mechanical vibration. When the excitation signal frequency is the same as the natural

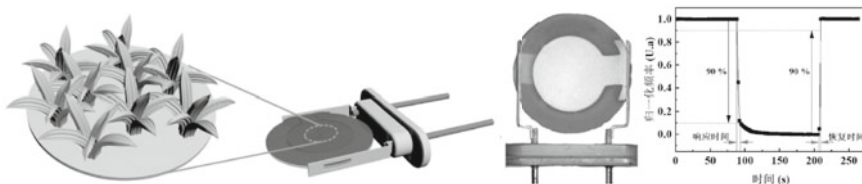


Fig. 7.68 QCM humidity sensor based on micro-nano structure of grass-like super-hydrophilic copper hydroxide

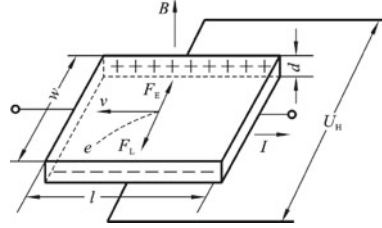
frequency of the wafer, resonance will occur, and the mechanical amplitude of the QCM will increase sharply. With the advantages of small size, simple structure, low energy consumption, high detection accuracy, low cost, high sensitivity, and online tracking, QCM has a wide range of applications in various fields such as chemistry, physics, biology, and materials, especially in gas sensors and humidity sensors. QCM itself does not have moisture-sensitive properties, it is not sensitive to liquids and humidity. Therefore, it is necessary to coat a moisture-sensitive film on the surface of the QCM electrode. The moisture-sensitive film will adsorb water molecules and cause the surface quality of the QCM to change, which in turn makes the QCM resonate. When the frequency is changed, the signal is obtained through certain technical means. Our team fabricated a highly sensitive humidity sensor with self-healing characteristics in water based on the QCM transducer and grass-like superhydrophilic copper hydroxide micro-nano structure. The sensitivity is as high as 85.9 Hz/RH%, and the response recovery time is up to 30 s/1.9 s. It can be effectively applied to detection applications such as sliding/moving of human fingers and breathing through mouth and nose (Fig. 7.68).

7.7 Hall Sensor

When a semiconductor with flowing electric current is exposed in magnetic field, a voltage difference is produced perpendicular to the current and magnetic field. This phenomenon is called Hall effect and it was discovered by Edwin Hall in 1879. Hall elements are usually made of InSb, GaAs, InAs and InAsP.

As shown in Fig. 7.69, with the presence of a magnetic field, electrons will experience Lorentz force and accumulate in one of the poles. Correspondingly, due to the lack of electrons, holes will accumulate in the other pole. The accumulation of electrons and holes gives rise to an electric field inside. Finally, the Lorentz force and the electric force will reach a balance, the electrons stop accumulating and a stable voltage is set up. Under the balance condition, the Lorentz force $F_L = ev \times B$ and electric force $F_E = eE = eU_H/w$ are equal. Then, we can obtain the Hall voltage:

$$U_H = vBw \quad (7.63)$$

Fig. 7.69 Hall effect

where v is the drift velocity for electrons, B is the magnetic field, w is the width of the element. The velocity is an unknown parameter for us, but we can find its relationship with current:

$$I = \frac{\Delta Q}{\Delta t} = \frac{\Delta t v w d n e}{\Delta t} = v w d n e \quad (7.64)$$

where d is the thickness of the element, n is the carrier concentration. Substituting Eq. (7.64) into Eq. (7.63), the Hall voltage can be written as:

$$U_H = \frac{1}{ne} \frac{IB}{d} = R_H \frac{IB}{d} \quad (7.65)$$

where R_H is the Hall coefficient.

The Hall voltage is proportional to the magnetic field, thus we can use the Hall element to measure magnetic field. However, the Hall voltage is small. The practical Hall sensor is an integrated chip with Hall element and amplifiers. There are mainly two configurations for the Hall integrated circuit: analog (linear) and digital (switch), as shown in Fig. 7.70. The linear sensor outputs a voltage that is proportional to the magnetic field, it is mainly used in measurement where the exact value of magnetic field is important. Whereas the Hall switch has an additional Schmitt trigger that binarize the signal with a threshold, it is mainly used in the situation where we do not care about the value of the magnetic field. The outputs of analog and digital Hall are displayed in Fig. 7.70c, d.

Example 7.26: Measure Rotational Speed with a Hall Switch Sensor

We have already shown several examples of measuring rotational speed with electromagnetic sensors in previous sections. The Hall sensor can also be used to measure rotational speed. The principle of measurement is shown in Fig. 7.71. As the notch or tooth passes by, the magnetic field in the vicinity of the Hall sensor is changed. With a Hall switch, we can obtain a signal of periodic pulses. By extracting the period of pulses, the rotational speed can be calculated.

Example 7.27: Current Clamp Based on Linear Hall Sensor

Current clamp is a device which allows the measurement of current in a wire without contact. The clamps can be open to let the wire in. According to Biot-Savart's law, the current in a straight wire generates magnetic field:

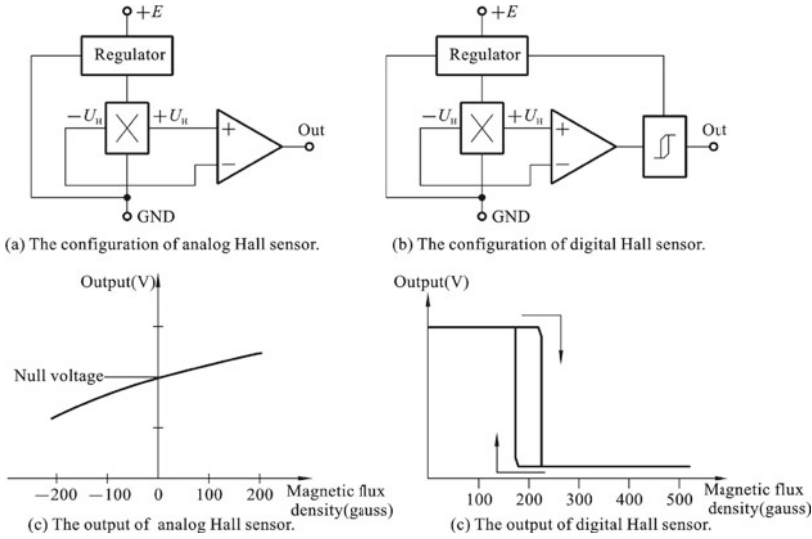
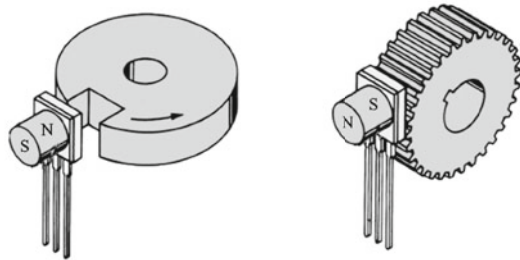


Fig. 7.70 Configurations and outputs of analog and digital Hall sensors

Fig. 7.71 Principle of measuring rotational speed with a Hall switch sensor



$$B = \frac{\mu_0}{2\pi} \frac{I}{r}$$

where r is the distance from the wire. Theoretically, we can calculate current by measuring the magnetic flux density at a single point and knowing the distance from that point to the wire. However, it is difficult to measure the distance precisely in practice. The current clamp solves this problem by considering the magnetic flux density in a loop. The clamps are made of ferromagnetic materials (ferrite or soft iron), so the magnetic field around the wire is concentrated in the loop of clamps. According to Ampere's law, the integration of magnetic flux density in any closed loop (regardless of shape) is related to the encircled current:

$$\oint B \cdot dl = \mu_0 I_{\text{encl}}$$



Fig. 7.72 Principle of current clamp

By inserting a linear Hall sensor in the loop of clamps to measure the magnetic flux density, the current flowing in the wire can be inferred (Fig. 7.72).

7.8 Photovoltaic Transducer

Photovoltaic effect refers to the generation of voltage in a material exposed to light. Photovoltaic effect is related to photoelectric effect since both of them state that the electrons are excited to a higher-energy state by absorbing the photon energy. However, there are some distinctions between them. Photoelectric effect is usually used to describe the phenomenon where electrons are ejected out of the material surface into vacuum, and photovoltaic effect describes the phenomenon where the electrons are still within the material.

Photovoltaic transducer is based on the photovoltaic effect and it is usually made of semiconductor p-n junctions. The commonly used materials include silicon (Si), germanium (Ge) and indium gallium arsenide (InGaAs). Its principle is shown in Fig. 7.73. When photons strike the semiconductor, electrons and holes are excited. Due to the built-in electric field in the depletion layer, electrons move to the n-type semiconductor while holes move to the p-type semiconductor. The respective accumulation of electrons and holes in the n-type and p-type semiconductors generates the voltage drop. If it is connected to a load, it can work as a battery.

Photovoltaic transducer mainly includes photovoltaic cell (i.e. solar cell), photodiode and phototransistor. Photovoltaic cell and photodiode are based on the same principle shown in Fig. 7.73. According to their usage, there are some special designs in the structure to enhance their certain properties. The photovoltaic cell is considered

Fig. 7.73 Principle of photovoltaic transducer

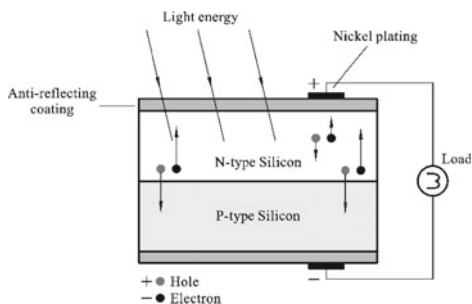
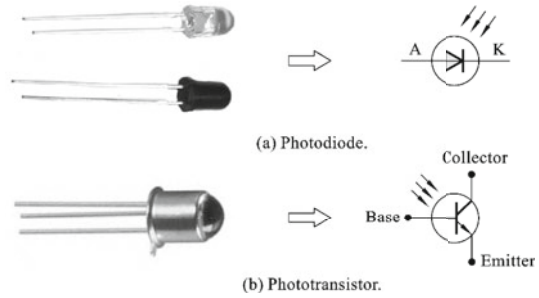


Fig. 7.74 Photodiode, phototransistor and their symbols



as a type of clean energy to provide electricity, hence it is designed to have a larger area and sensitive bandwidth to maximize the energy to be converted. Whereas the photodiode is used as a sensor, hence the sensitivity and speed is the main properties to be optimized. The phototransistor is made of PNP or NPN type semiconductors. It provides an additional gain to the photodiode, resulting in a much larger sensitivity (Fig. 7.74).

Photodiode and photoresistor are both light sensitive sensors, however they are different types of sensors. The photodiode is an energy conversion type sensor which directly converts the photon energy into electric energy while the photoresistor is an energy control type sensor which requires the assistance of a power supply to convert the change of light intensity into the change of voltage.

Example 7.28: Photoelectric Switch

We have introduced some applications of proximity switches and how to make them based on electromagnetic sensors in the previous sections. If the tested object is a non-ferromagnetic insulator, the photoelectric proximity switch is more suitable. Its working principle is shown in Fig. 7.75. An emitter sends a light (usually infrared) to the receiver which consists photodiodes or phototransistors. When an object passes by in between the emitter and receiver, it blocks some of the light and reduces the photon energy that hits the receiver. By checking the change of the output signal of the photodiodes or phototransistors, the presence of an object can be determined.

Example 7.29: Perovskite Photodetector

Currently, commercial photodetectors are mainly semiconductors based on inorganic compounds of Si, GaN and InGaAs. In the manufacturing of these device, high-vacuum and high-temperature environment is required, which makes them expensive. In addition, an additional bias voltage is required during the use of the device to



Fig. 7.75 Principle of photoelectric proximity switch

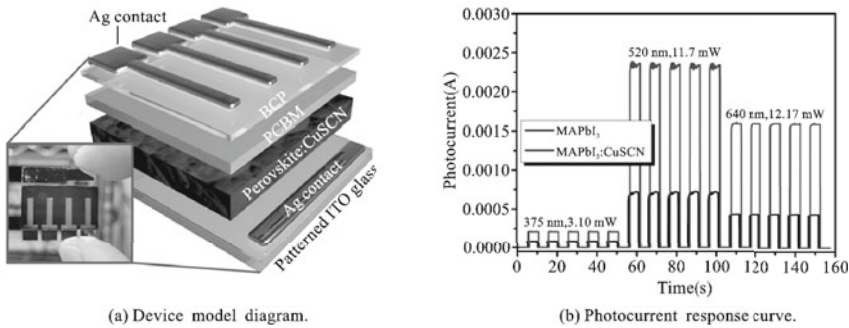


Fig. 7.76 Perovskite photodetector

obtain a better gain, which increases energy consumption. Therefore, searching for low-cost, self-driving new semiconductor materials to replace traditional inorganic semiconductors is an important field of photoelectric detection. Organic–inorganic hybrid perovskite materials have the advantages of low cost, low temperature solution preparation, high light absorption coefficient, adjustable band gap, long diffusion length and fast charge transfer, etc., and have been greatly developed in the field of photoelectric detection. Our team has proposed a new type of self-driving photodetector based on perovskite materials without a hole transportation layer. And it has been successfully applied to the visible light communication system. The accurate transmission of text and the high-fidelity conversion of audio are realized (Fig. 7.76).

DIY Experiment 7.8 Automatic Line Tracking Robot Car

This is a challenging experiment that requires not only the knowledge of sensors and Arduino but also the knowledge of mechanical design. The designed robot car is supposed to be able to track and move along a black line drawn on a white board. A module consists of four LED-photodiode pairs is used. When a LED-photodiode pair is above the black line, the photodiode cannot receive reflected light due to the absorption of the black line. Thus we can determine and change the orientation of the robot car by the output signals in the photodiodes. For example, if all the three photodiodes receive reflected light while the rightmost one does not, it means the robot car is on the left-hand side of the black line and the Arduino board should send a control signal to turn the robot car to the right (Fig. 7.77).

7.9 Image Sensors

Image sensor converts optical images into electrical signals. There are two types image sensors: CCD (charge coupled device) and CMOS (complementary metal oxide semiconductor). CMOS has low power consumption, low cost, and high speed, while CCD has the mature technology, less image noise, and high imaging quality.

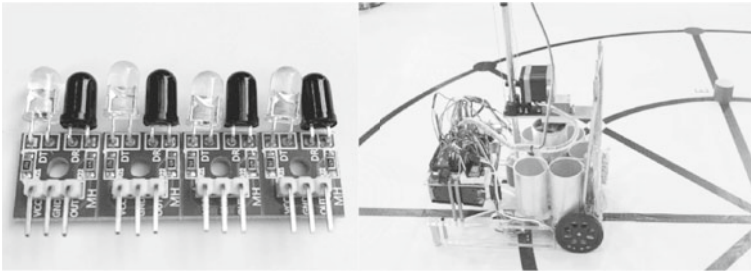


Fig. 7.77 Line tracking robot car

The image sensor can be regarded as an optical sensor array with position information. Image sensors have been widely used in digital cameras, scanners, object detection, defect inspection and industrial control.

A CCD consists of an array of coupled p-doped MOS (metal oxide semiconductor) capacitors. The p-doped MOS capacitor is the elementary building block for the CCD, which is made by growing a layer of silicon dioxide (SiO_2) on top of a p-type silicon substrate, and depositing a layer of metal or polycrystalline silicon above it. The three layered structure is equivalent to a parallel plate capacitor, with one of the electrodes replaced by a semiconductor.

When a positive bias voltage is applied to the metal electrode (gate), the majority carriers (holes) in the silicon substrate are pushed to the bottom of the substrate which leaves a layer of immobile acceptors ions. Thus, a depletion layer is formed in the Si/SiO₂ interface. As the bias voltage increases, the depletion layer becomes deeper, and if the bias voltage is higher than a threshold value, the minority carriers (electrons) are attracted to the region near the interface and create a very thin inversion layer with high density of electrons. The deeper the depletion layer the higher the silicon's ability to attract electrons. Usually, "potential well" is used to describe the ability of MOS attracting electrons. We can consider the potential well as a bucket and the electrons as raindrops, each bucket has its limit to collect raindrops (Fig. 7.78).

A MOS is also sensitivity to light. Recall that, in Sect. 7.8, we mentioned that silicon is a suitable material to make photodiode. Thus, when the shutter opens, free

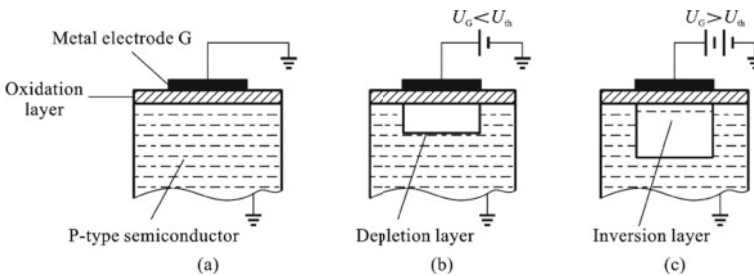


Fig. 7.78 Configuration of a MOS and its depletion layer

electrons are created in the silicon substrate of the MOS. In another word, the MOS has both the abilities of creating and collecting electrons when exposed to light.

The final task for the CCD is the read the photon induced charges in each pixel when the exposure is finished (after the shutter is closed). A technique similar to the shift register is used to read the induced charges sequentially, the process is shown in Fig. 7.79. It shows a one-dimensional CCD with only four pixels. Each time, the charge on the rightmost pixel is read using a charge to voltage converter (CVC) and an analog to digital converter (ADC), and the charges in other pixels keep transferring to its neighboring pixel on the right. After the charges in all pixels are read, a digital image can be displayed.

The transferring of charge is done by changing the voltage applied to each MOS and the corresponding potential well. The process is illustrated in Fig. 7.80. In this example, each pixel has three MOS capacitors to collect charges. The voltage sequences in Fig. 7.80a are applied to the electrodes in Fig. 7.75b. As already explained, the larger the applied voltage, the deeper the depletion layer and the potential well. Therefore, the electrons in one MOS are attracted by its neighboring MOS if the voltage applied to the neighboring one is higher. This process is like pouring the water in one bucket to its neighboring bucket one by one. The distribution of charges at some selected time instances are displayed in Fig. 7.80c.

For the CMOS image sensor, each pixel has its own photodetector and amplifiers. Thus, the CMOS sensor is regarded as active-pixel sensor. Usually, each pixel is consisted of a pinned photodiode, a floating diffusion and four CMOS transistors (a transfer gate, a reset gate, a selection gate and a source-follower readout transistor). A column driver and a row driver are used to read output of each pixel. A comparison of CCD and CMOS is made in Table 7.4.

The image sensors, either CCD or CMOS, can only sense the intensity of light without color information. Therefore, when camera was invented in the early times, it can only take black-and-white photos. To get a color image, filters are placed in front of the photodiodes as shown in Fig. 7.81, and the pixels are divided into groups of four pixels. For each group, two pixels are sensitive to green, one pixel is sensitive to red and one pixel is sensitive to blue. Therefore, three filtered images in red, green and blue are obtained. Finally, we can add these filtered images to reconstruct the

Fig. 7.79 Process of reading charges sequentially

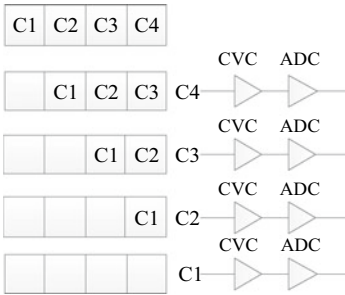


Fig. 7.80 Principle of charge transfer

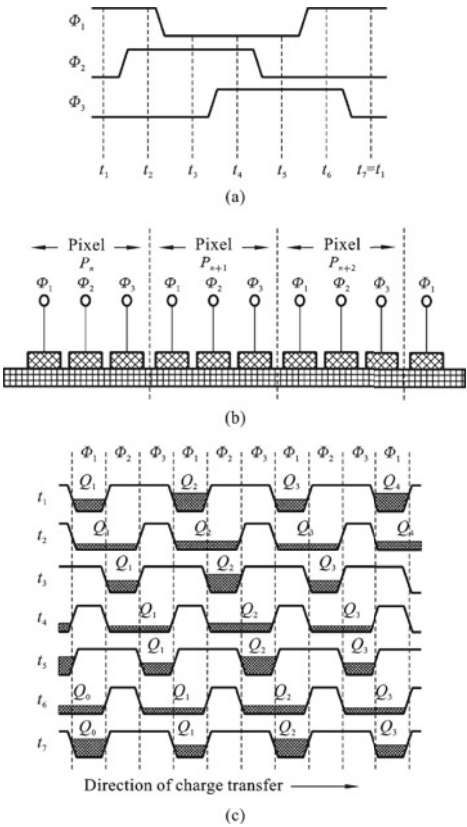


Table 7.4 Comparison of CCD and CMOS

	CCD	CMOS
Cost	High	Low
Sensitivity	High	Low
Power consumption	High	Low
Voltage	12 V	5 V or 3.3 V
Size	Large	Small

original color image. The image is also called a RGB image because of the colors we used in the filtering and reconstruction.

Example 7.30: Barcode Scanner

Barcode is a way of representing data by parallel lines with different widths and spacing. Barcodes have been printed in books and a lot of glossary items. The bookstore and supermarket usually save a barcode database for all the items that have been stocked. Thus, a cashier only needs to scan the barcode to get the price of the

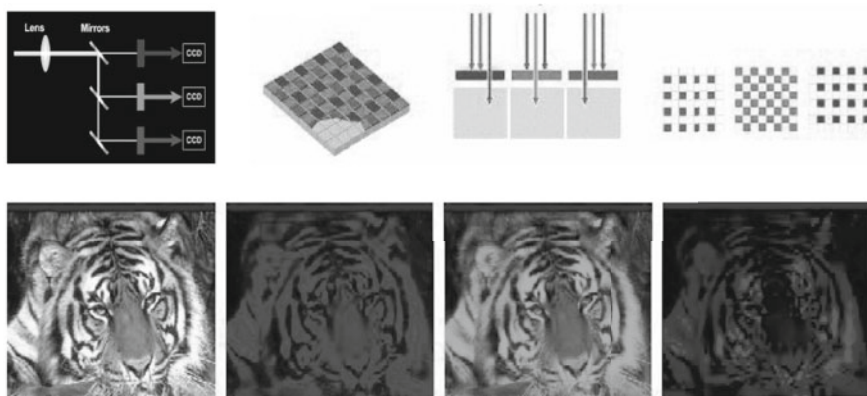


Fig. 7.81 Principles of RGB image

item, which greatly facilitates the process of payment. The barcode scanner contains a linear array of CCDs which detects the light reflected from the barcode. Then, the signal is converted into binary digits and compared with database to find the item. Nowadays, the usage of QR code (Quick Response code), which is a two-dimensional barcode, is booming, especially in China. It has been used for payment, website login, restaurant ordering, ticket displaying, Wi-Fi connection and so on. And we can use the camera of our smart phone to scan it directly (Fig. 7.82).

Example 7.31: Inspection of Bearing Defect Based on Machine Vision

Machine vision is a technology that mimics human vision, in which CCDs or CMOSs are used to obtain digital images and computers are used to analyze the images. Machine vision is replacing human vision in some fields such as defect inspection. Traditionally, the defects in bearings are inspected by workers visually. However, bearings are usually mass-produced, so the workers need to examine many bearing each day. The inspection result highly depends on the experience of worker and it is easily influenced by the tiredness of workers. With machine vision, the defects can be automatically found by analyzing the images with a software.

Fig. 7.82 Barcode and QR code



7.10 Thermocouple

Thermocouple is a temperature measuring device consisting of two dissimilar electrical conductors jointed at both ends to form a loop as shown in Fig. 7.83. Due to the thermoelectric effect, when the temperatures are difference at the two junctions, an electromotive force is induced in the loop. This phenomenon was first discovered by German physicist Thomas Johann Seebeck in 1821, and it is called Seebeck effect. Thermocouple only measures the temperature difference between two junctions, hence it is a relative temperature measuring sensor. To measure the absolute temperature, one of the junctions should be placed in an environment with known temperature, such as the ice-water mixture of 0 °C or boiling water of 100 °C. The junction that is used as the reference is called a cold junction or reference junction, while the other one is called hot junction or working junction.

7.10.1 Thermoelectric Effects

1. Seebeck effect

In 1821, Seebeck discovered that, when there is a temperature difference at the junctions of two metals, a nearby compass shows a deflection of the needle. He referred it as thermomagnetic effect. But, afterwards, the induced magnetic field is found to be originated from the thermo-electric current in the conductors. He also tested different combinations of metals and found that the current intensity and direction is related to the metals that have been used.

Seebeck effect is mathematically expressed as:

$$V_S = \int_{T_2}^{T_1} (\alpha_A - \alpha_B) dT \quad (7.66)$$

where V_S is the induced voltage due to Seebeck effect, α_A , α_B are the Seebeck coefficients of metal A and B, T_1 and T_2 are the temperatures at the two junctions.

2. Peltier effect

Peltier effect refers to the phenomenon that when current is flowing in a loop made of two metals, heat is absorbed in one junction while released in the other one, as shown

Fig. 7.83 Configuration of a thermocouple

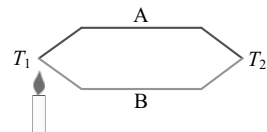
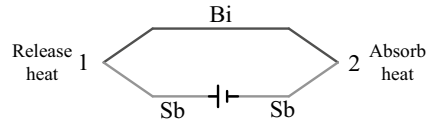


Fig. 7.84 Schematic of Peltier effect



in Fig. 7.84. It was discovered by French physicist Jean Charles Athanase Peltier in 1834. It should be noted that, although this heat is also associated with current, it is obviously different from the Joule heat because it is related to the direction of current. When the direction of power supply changes in Fig. 7.84, Junction 1 absorbs heat while Junction 2 releases heat.

Peltier effect can be explained by contact potential. When two dissimilar metals are brought into contact, electrons will diffuse from one material to the other due to the difference in electron density. The metal that loses electrons is positively charged at the interface, while the metal receives electrons is negatively charged. The charged thin layer creates a potential that hinders the movement of electrons and finally an equilibrium is reached with a fixed contact potential. The contact potential is related to the electron density of the metals and the temperature:

$$E_{AB}(T) = \frac{kT}{e} \ln \frac{N_A}{N_B} \quad (7.67)$$

where E_{AB} is contact potential, k is the Boltzmann constant, e is the elementary charge, N_A and N_B are electron density for metal A and B. Considering the contact potential as a battery, then the battery in Junction 1 is under charging, the non-electrostatic force does negative work and Peltier heat is released. The battery in Junction 2 is discharging and absorbs heat.

The total electromotive force due to the contact potential is:

$$E_P(T_1, T_2) = E_{AB}(T_1) - E_{AB}(T_2) \quad (7.68)$$

It can be noticed from Eq. (7.68) that, if the temperatures at two junctions are the same, then there is no electromotive force in the loop.

3. Thompson effect

Although the contact potential relates the electromotive force with temperature, people found that the equation does not satisfy the experimental result. Then, William Thompson (Lord Kelvin) stated that, the temperature gradient inside each metal also induces an electromotive force. Electrons in the region with higher temperature possess higher kinetic energy and diffuse into the region with lower temperature. This process is equivalent to applying a non-electrostatic force to the electrons. The equivalent electric field for the non-electrostatic force is proportional to the gradient of temperature: $\sigma \cdot dT/dl$, where σ is the Thompson's coefficient. Thompson's coefficient is positive for some materials such as Cd, Zn, Ag, Cu and it is negative for some materials such as Fe, Pt, Pd. The induced electromotive force in one metal due

to Thompson effect is:

$$E_T = \int_0^l \sigma \frac{dT}{dl} dl = \int_{T_2}^{T_1} \sigma dT \quad (7.69)$$

Taking both Peltier and Thompson effects into account, Seebeck's experiment is explained. The total thermoelectric electromotive force in the thermocouple shown in Fig. 7.83 is:

$$E_{AB}(T_1, T_2) = E_{AB}(T_1) - E_{AB}(T_2) + \int_{T_2}^{T_1} \sigma_B dT - \int_{T_2}^{T_1} \sigma_A dT \quad (7.70)$$

The electromotive force in Eq. (7.70) is equal to the net Seebeck electromotive force in Eq. (7.66).

7.10.2 Thermoelectric Laws

Law No. 1 Nonhomogeneous material law

This law states that a thermocouple must be made with two or more dissimilar materials. If the thermocouple is made of homogeneous material, it can be easily noticed from Eq. (7.70) that the thermoelectric potential is zero regardless of the temperature difference between two junctions.

Law No. 2 Intermediate temperature law

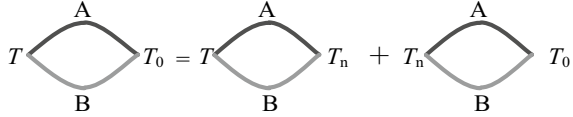
For a thermocouple, if the thermoelectric potentials for temperature pairs (T, T_n) and (T_n, T_0) are known, then the thermoelectric potential for temperature pair (T, T_0) can be directly calculated:

$$E_{AB}(T, T_0) = E_{AB}(T, T_n) + E_{AB}(T_n, T_0) \quad (7.71)$$

where T_n is an intermediate temperature. Equation (7.71) can be easily proved by Eq. (7.70). This law indicates that we can extend a thermocouple made of metals A and B with the same materials, and the thermoelectric potential only depends on the temperatures at two ends, and is irrelevant to the intermediate temperature as shown in Fig. 7.85.

The intermediate temperature law also enables measurements with different reference temperature. When using thermocouple, it is quite difficult to calculate temperature from measured thermoelectric potential using Eq. (7.70). Thus, thermocouples are calibrated before they are put into use. During the calibration, the cold junction is placed in an environment with fixed temperature (e.g. 0 °C), and thermoelectric

Fig. 7.85 Demonstration of the intermediate temperature law



potentials are measured with hot junction placed in different temperatures. A table corresponds temperature and thermoelectric potential is obtained after the calibration. During the measurement, the table is used to find the temperature corresponds to the measured thermoelectric potential. Although the table is obtained with a reference temperature of 0 °C, it can also be used for other reference temperatures by taking 0 °C as the intermediate temperature in Eq. (7.71).

Law No. 3 Intermediate conductor law

Inserting a third material to the thermocouple will not change the induced thermoelectric potential, as long as the inserted material has the same temperature at both ends. This law can be mathematically proved by the following equation, with the configuration shown in Fig. 7.86.

$$\begin{aligned}
 E_{ABC}(T, T_0) &= E_{AB}(T) + E_{BC}(T_0) + E_{CA}(T_0) + \int_T^{T_0} \sigma_A dT + \int_{T_0}^{T_0} \sigma_C dT + \int_{T_0}^T \sigma_B dT \\
 &= \frac{kT}{e} \ln \frac{N_A}{N_B} + \frac{kT_0}{e} \ln \frac{N_B}{N_C} + \frac{kT_0}{e} \ln \frac{N_C}{N_A} + \int_T^{T_0} \sigma_A dT + \int_{T_0}^{T_0} \sigma_C dT + \int_{T_0}^T \sigma_B dT \\
 &= \frac{kT}{e} \ln \frac{N_A}{N_B} + \frac{kT_0}{e} \ln \frac{N_B}{N_A} + \int_T^{T_0} \sigma_A dT + \int_{T_0}^T \sigma_B dT \\
 &= \frac{kT}{e} \ln \frac{N_A}{N_B} - \frac{kT_0}{e} \ln \frac{N_A}{N_B} + \int_{T_0}^T \sigma_B dT - \int_{T_0}^T \sigma_A dT \\
 &= E_{AB}(T, T_0)
 \end{aligned}$$

This intermediate conductor law makes thermocouple a practical sensor, it justifies the inserting of a measurement circuit to measure the induced thermoelectric potential as shown in Fig. 7.87. It also indicates that there is no need for the measuring instrument to be around the thermocouple, we can connect a compensation wire to the thermocouple and make the measurement far away. It should be mentioned that, although the intermediate conductor law indicates any material could be used

Fig. 7.86 Demonstration of the intermediate conductor law

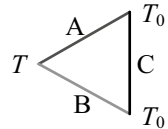


Fig. 7.87 Measuring thermoelectric potential with compensation wire

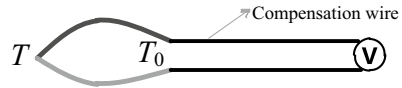
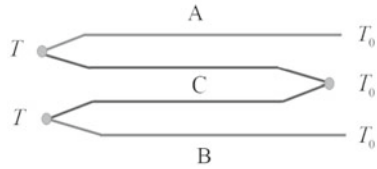


Fig. 7.88 Demonstration of the reference electrode law



to make the compensation wire and would not change the induced thermoelectric potential, it is better to use a material with similar property as the thermocouple to make the compensation wire to avoid errors caused by the fluctuation at the ends.

Law No. 4 Reference electrode law

If a thermocouple is made of two materials that are not commonly used, the thermoelectric potential is unknown before it is calibrated. However, we can insert a commonly used material C as show in Fig. 7.88 to form two thermocouples. Then, the induced thermoelectric potential of the original thermocouple is the sum of the two thermocouples:

$$E_{AB}(T, T_0) = E_{AC}(T, T_0) + E_{CB}(T, T_0)$$

7.10.2.1 Types of Thermocouple

Thermocouples can be made of different combinations of materials. They differ in sensitivity, range, linearity and cost. According to the application, thermocouple type should be selected properly. Some commonly used types are listed in Table 7.5.

The characteristic curves for some selected thermocouples are depicted in Fig. 7.89.

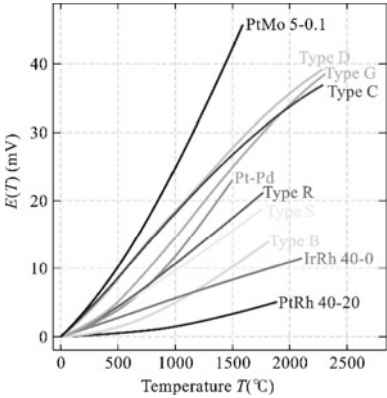
7.10.2.2 Characteristics and Applications of Thermocouple

Thermocouple differs from other temperature sensors (thermistor and resistance thermometer) that have been introduced in the following aspects: (1) Thermocouple is a passive sensor, which converts the heat energy directly into electric energy while thermistor and resistance thermometer are active sensors that require an external power supply; (2) Thermocouple is a relative temperature sensor which measures the

Table 7.5 Common types of thermocouples

Type		Material	Temperature range	
			Long term	Short term
Nickel-alloy	Type E	Chromel–constantan	0–800 °C	–40 to 900 °C
	Type J	Iron–constantan	0–750 °C	–180 to 800 °C
	Type K	Chromel–alumel	0–1100 °C	–180 to 1370 °C
	Type N	Nicrosil–Nisil	0–1100 °C	–270 to 1300 °C
	Type T	Copper–constantan	–185 to 300 °C	–250 to 400 °C
Platinum/rhodium-alloy	Type B	70%Pt/30%Rh–94%Pt/6%Rh	200–1700 °C	0–1820 °C
	Type R	87%Pt/13%Rh–Pt	0–1600 °C	–50 to 1700 °C
	Type S	90%Pt/10%Rh–Pt	0–1600 °C	–50 to 1750 °C
Tungsten/rhenium-alloy	Type C	95%W/5%Re–74%W/26%Re	N/A	N/A
	Type D	97%W/3%Re–75%W/25%Re	N/A	N/A
	Type G	W–74%W/26%Re	N/A	N/A

Fig. 7.89 Characteristic curves for selected thermocouples



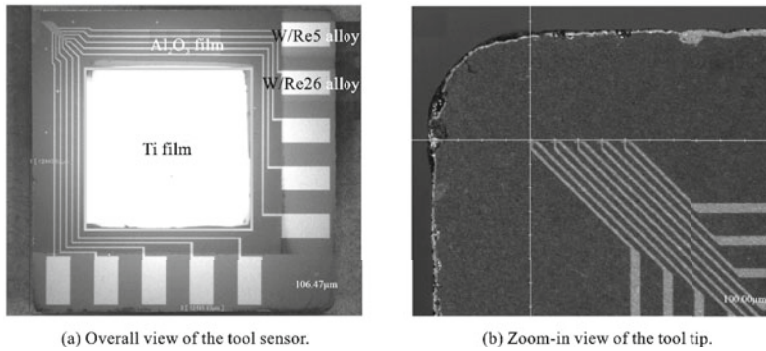


Fig. 7.90 Tool temperature sensor based on thin film thermocouple

temperature difference at two junctions. To measure an absolute temperature, a reference temperature with known value is required. Whereas, thermistor and resistance thermometer measure absolute temperature directly.

Besides, thermocouple has larger measuring range than thermistor and resistance thermometer. It is suitable for temperatures higher than 1000 °C. Thus, it is commonly used high temperature measuring cases such as monitoring the temperature of a furnace.

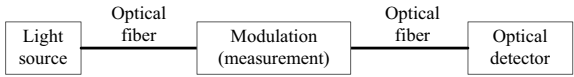
Example 7.32: Tool Temperature Sensor Based on Thin Film Thermocouple

In the cutting process, the cutting temperature and its distribution directly affect the machining quality of the workpiece and the service life of the tool. However, due to the harsh cutting environment (large strain/stress, steep temperature gradient) in the narrow cutting area, it is difficult for conventional temperature sensor to approach the cutting area and realize the in-situ temperature measurement of the cutting area. Our team used MEMS technology to fabricate the thin-film thermocouple array on the tool substrate (Fig. 7.90). Combined with an improved hot-press diffusion welding process, the sensor array is integrated in the tool body. Thus the in-situ measurement of the cutting temperature of tool tip is realized.

7.11 Fiber-Optic Sensor

Optical fiber is a flexible and transparent wire made of silica or plastic. It has a diameter similar to human hair. Fiber-optic sensor is based on optical fiber, it can be used to sense any physical quantity that interacts with the light transmitting in the fiber. The major advantage of fiber-optic sensor is the multiplexing, which means that signals with different frequencies can transmit in the same optical fiber together. Therefore, a single fiber-optic sensor can be used to measure several physical quantities, such as pressure, strain and temperature, at the same time, with one bandwidth allocated for one physical quantity. In addition, because fiber-optic sensor uses light

Fig. 7.91 Signal transmission process in a system with optical fiber



to carry information, it has relatively good resistance to electromagnetic noises and can work under harsh conditions with strong electromagnetic interference.

The working principle of the fiber-optic sensor is shown in Fig. 7.91. The light generated by the light source is directed into the optical fiber and transmits within it. The physical quantity to be measured interacts with optical fiber at some point and changes some properties (such as intensity, frequency/wavelength, phase, polarization state, etc.) of the light. This interaction is similar to the process of modulation where the light is a carrier signal. The modulated signal keeps transmitting in the optical fiber and is demodulated at the optical detector.

According to the way of modulation, fiber-optic sensors can be mainly classified as frequency/wavelength type and light intensity type.

7.11.1 Frequency/Wavelength Type

The frequency/wavelength type is more commonly used, and it is also named fiber Bragg grating (FBG). Fiber Bragg grating is an optical fiber with periodic change of refractive index as shown in Fig. 7.92. Due to the Fresnel reflection, when broadband lights encounter a grating, the light possessing a particular wavelength is reflected while others pass without interference. The reflected wavelength, which is also called Bragg wavelength is given by:

$$\lambda_B = 2n_c \Lambda \tag{7.72}$$

Fig. 7.92 Structure and working principle of fiber Bragg grating

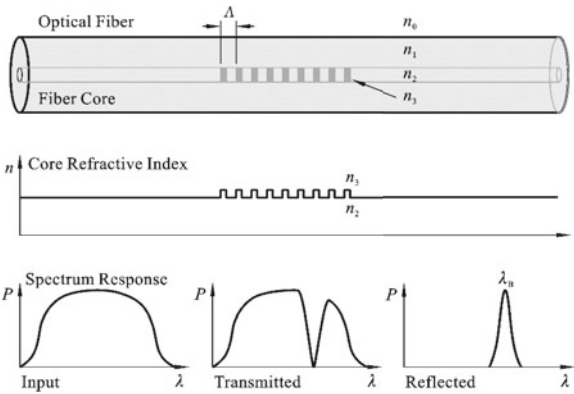




Fig. 7.93 Mechanical change of fiber Bragg grating under stress

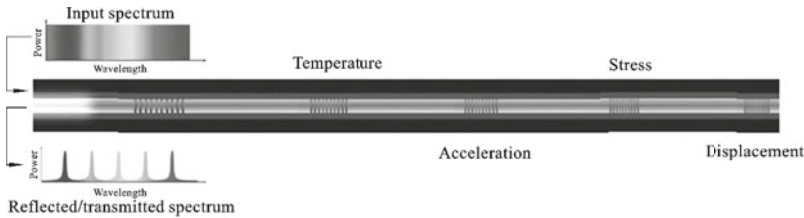


Fig. 7.94 Illustration of multiplexing

where λ_B is the Bragg wavelength, n_e is the effective refractive index, Λ is the grating period.

When a fiber Bragg grating sensor is stressed, the optical fiber expands and causes the change of grating period as shown in Fig. 7.93. The change of grating structure further causes a shift in the reflected and transmitted spectrum. Therefore, by measuring the change of reflected or transmitted frequency, the stress can be inferred. Similarly, other physical quantities such as temperature, acceleration and displacement also change the grating period by deforming the optical fiber, so they can also be measured by the fiber-optic sensor.

A fiber Bragg grating sensor can measure multiple physical quantities and carry multiple types of information at the same time. This characteristic is called multiplexing and it is achieved by making multiple gratings with different period in the same optical fiber as shown in Fig. 7.94. Because there are several gratings in the same optical fiber, the reflected spectrum has multiple peaks as shown in Fig. 7.94. Usually, the measured physical quantity only changes Bragg wavelength in a small interval. Therefore, even there are multiple peaks in the same spectrum, they will not interfere with each other, and we can estimate each physical quantity by the shift of corresponding peak in the spectrum.

7.11.2 Intensity Type

The light intensity can also interact with other physical quantities, and can be used to carry information. An example of the intensity type fiber-optic sensor is shown in Fig. 7.95. This sensor can be used to measure temperature. The metal sheet deforms with the increase of temperature and applies a pressure to the clamp. The clamp

further bends the optical fiber and causes the partial reflection of light. Thus, the light intensity that is measured at the other end is related to the temperature.

Example 7.33: Measuring Wing Deformation by Fiber-Optic Sensor

When a drone is flying, the deformations of the wings should be monitored to ensure safety. Traditionally, strain gauge networks are attached to wings to monitor strains in multiple locations. The wings have a very large area, thus it requires lots of strain gauges and an instrument with large number of channels to process the data. The large number of strain gauges brings the drawback of weight increase. With the development of optical fiber, the fiber-optic sensor is replacing strain gauges in measuring deformation. As illustrated in Fig. 7.96, a single optical fiber can measure the strains at multiple locations along the fiber path. The use fiber-optic sensor reduces cost of instrument by reducing the number of channels and also reduces the weight of the drone.

Example 7.34: On-line Monitoring of Cutting Temperature Based on Micro-probe Multi-band Mid-infrared Optical Fiber Sensing

In metal cutting, the cutting temperature directly affects the surface integrity, machining accuracy and tool life of the part. However, the in-situ online measurement of cutting tool temperature faces a huge challenge due to strong interference such as strong time-varying local high temperature in the tool tip, large temperature gradient, as well as cutting fluid. With the merits of small size, high temperature resistance, anti-electromagnetic interference, high chemical stability of the optical fiber, it is integrated into the tool to realize the non-contact in-situ online measurement of the cutting temperature of the tool. Our team has developed a cutting temperature online measurement system based on micro-probe multi-band mid-infrared

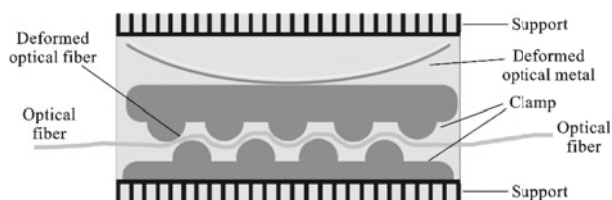


Fig. 7.95 Fiber-optic temperature sensor



Fig. 7.96 Measuring wing deformation with strain gauges and optical fiber

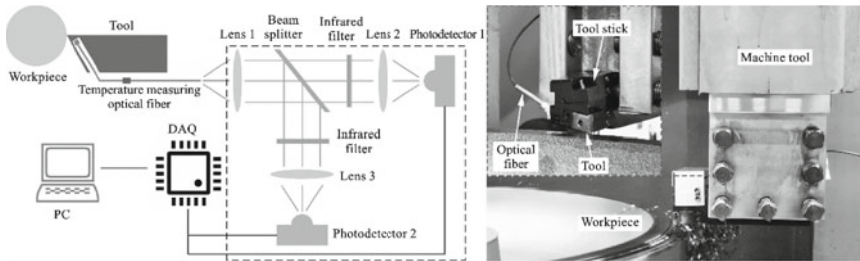


Fig. 7.97 On-line monitoring system for cutting temperature based on micro-probe multi-band mid-infrared optical fiber sensing

fiber optic sensor. It can monitor temperature up to 1500 °C with an accuracy of 1 °C (Fig. 7.97).

Example 7.35: Monitoring Pipeline Leakage

In Example 7.25, we introduced using ultrasonic flowmeter to monitor pipeline leakage. Pipelines are used to transmit gas and oil between cities or even countries, thus their length can reach thousands of kilometers. Ultrasonic flowmeters need to be installed at intervals to monitor the whole pipeline. Only one optical fiber is need to solve the problem. An optical fiber with multiple gratings can be embedded into the pipeline to monitoring any pressure change caused by the leakage. At the same time, it can also monitor other important parameters such as temperature and vibration.

7.12 Grating Sensor

Grating is an optical device consisting dense and equidistant parallel slits. A grating sensor is consisted of a light source, a photodetector and two gratings (a scale grating and an indicating grating) as shown in Fig. 7.98. Usually, the indicating grating is shorter and has the same grating period (the distance between the center of a slit to the center of an adjacent one) as the scale grating. During the measurement, the scale grating is stationary while the indicating grating moves with the measured object.

When two gratings are overlapped with an angle, moiré fringe (also known as moiré pattern) is formed as illustrated in Fig. 7.99. Moiré fringe is the visual effect of the interference between the parallel slits. The width of moiré fringe is given by:

$$W_M = \frac{\Lambda}{2 \sin(\theta/2)} \approx \frac{\Lambda}{\theta} \quad (7.73)$$

where W_M is the width of moiré fringe, Λ is the grating period, θ is the angle between the two gratings. The approximation in Eq. (7.73) holds when θ is small. As shown in Fig. 7.99, when the angle between two gratings changes, the width of moiré fringe

Fig. 7.98 Principle of grating sensor

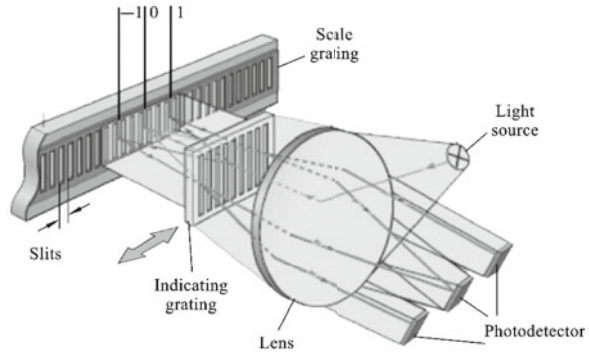
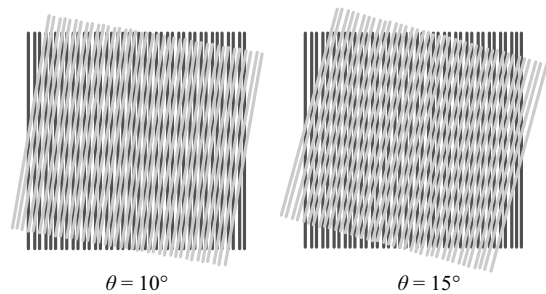


Fig. 7.99 Moiré fringes formed by two overlapped gratings with different angles



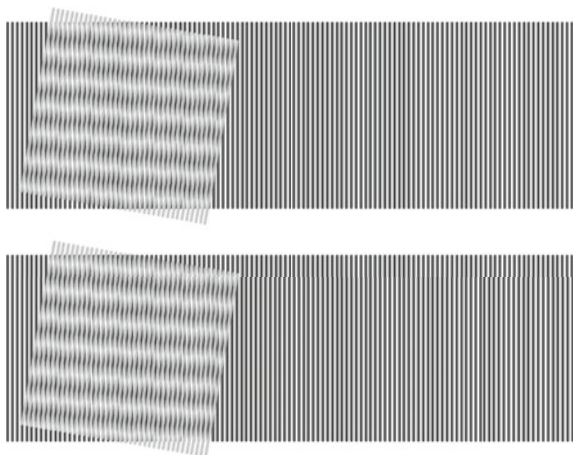
changes accordingly. If a grating is connected to a shaft, a small rotation of the shaft can be detected by measuring the change of fringe width.

The grating sensor can also be used to measure linear displacement. As shown in Fig. 7.100, the linear movement of the top grating causes a relatively large movement of moiré fringe. According to Eq. (7.73), the fringe width is much larger than the grating period if the angle between two gratings is small. Measuring the movement moiré fringe is much easier than measuring the tiny movement of a grating. Thus, the grating sensor acts like an optical amplifier to improve the measurement accuracy. This effect can be observed in Fig. 7.100, where a small movement of fringe can be detected while the movement of grating is difficult to observe.

Example 7.36: Precision Displacement Measurement of Drilling Machine Tools

In order to ensure machining accuracy, the movement of the drill or tool must be accurately monitored. Linear gratings can be used to measure the displacement of drills or tools. The moiré effect can amplify tiny movements, so accurate displacement measurement can be achieved.

Fig. 7.100 Change of moiré fringes due to movement of a grating



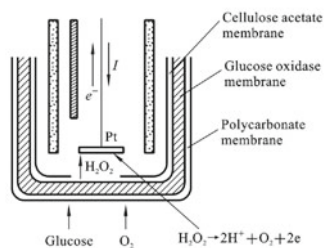
7.13 Biosensor

Biosensor is a device integrating a biological element with a conventional physical sensor to convert the biochemical quantities into physical quantities such as heat and light, then further into electrical signals. Biosensors are developing very fast, and they have been widely used in medical instruments to monitor many indexes of human body such as blood glucose, blood oxygen, lactic acid, etc. According to the interaction principle of biological element, biosensor can be classified into enzymatic sensor, immunosensor, microorganism sensor, etc.

7.13.1 Enzymatic Sensor

Enzymes are proteins produced by organisms which have catalytic ability. They catalyze the chemical reactions of some particular molecules. A basic step in the process of catalysis is to combine the substrate with the enzyme and convert it into another chemical product. Therefore, the enzyme has the dual functions of molecular recognition and transformation. As a biocatalyst, enzyme has high degree of specificity compared with general catalyst. An enzyme can only act on a certain kind of substrate. The enzyme has very high efficiency of catalysis, each enzyme molecule converts 10^3 – 10^6 substrate molecule per minute. Since enzymes are proteins, they will be inactivated in high temperature and acid environments, the

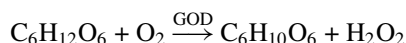
Fig. 7.101 Configuration of a glucose meter



catalysis must undergo in an appropriate environment. Besides, enzymes are water-soluble substances, so they cannot be used in sensor directly. An appropriate carrier is needed to form a water-insoluble layer for the enzyme.

Example 7.37: Blood Glucose Meter

Glucose meter is a device to measure the concentration of glucose. It can be used in blood glucose monitoring for diabetic patients. A typical glucose sensor is composed of three layers of membranes. The inner membrane layer serves as a substrate to immobilizing enzymes, and is also a $\text{H}_2\text{O}_2/\text{O}_2$ selection membrane. The middle layer is where the chemical reaction happens, which converts glucose into H_2O_2 . The outer layer is used to control the transfer of glucose and make the reaction under control. The chemical reaction of glucose catalyzed by glucose oxidase (GOD) is:



To measure the concentration of glucose, we can either measure the consumption of oxygen or the yield of hydrogen peroxide. The configuration shown in Fig. 7.101 is to measure the yield of hydrogen peroxide. The glucose and oxygen in the sample pass through the polycarbonate membrane and penetrate into the fixed GOD membrane. Then, the generated H_2O_2 pass the selective membrane and is oxidized on the platinum anode and generates an electric current. By measuring the generated current, the glucose concentration can be inferred.

7.13.2 Microorganism Sensor

To solve the problem of lacking enzyme source and reduce the cost (enzyme is expensive) of sensor, some new biosensors are developed, such as the microorganism sensor. A typical microorganism sensor is based on principles similar to the enzyme sensor. The difference comes from the use of microorganism. Microorganism sensor can be divided into current type and potential type. Generally speaking, the current type is superior to the potential type in the following aspects: its output signal is directly proportional to the concentration of the measured substance; the relative

error of the concentration corresponding to the reading error is small; the sensitivity is higher.

The microorganism sensor can be divided into two types according to microorganisms used:

- (1) For aerobic microorganisms, while interacting with the substrate (called assimilated organic matter), the respiration activity of their cells increases and the oxygen consumption increases. By using oxygen electrode or CO electrode to measure its respiratory activity, the substrate concentration can be calculated. This type of sensor is respiratory activity measurement type.
- (2) For anaerobic microorganisms, after the microorganisms assimilate the tested organic matter, various metabolites, such as CO_2 , H_2 , H_+ , etc., are generated. The concentration of metabolite can be measured to infer the substrate concentration. This type of sensor is a metabolite measuring type sensor.

Figure 7.102 shows the structure of these two types of microorganism sensor. In Fig. 7.102a, the aerobic microorganism membrane is installed on the Clark oxygen electrode. Inserting the electrode into a sample containing organics that can be assimilated, the organics diffuse to the microorganism membrane so that they are assimilated by the microorganism. Thus, the amount of oxygen diffused on the oxygen probe is reduced, and the correspondingly current of the oxygen electrode decreases. The organics that have been assimilated can be obtained by measuring the current. In Fig. 7.102b, the bacteria which produces H_2 is fixed to a membrane which is installed on the Pt anode of the cell. Ag_2O_2 is used as the cathode and phosphate as the electrolyte. When the sensor is inserted into a solution containing organics, the organics are assimilated by the H_2 -producing bacteria to produce H_2 . The produced H_2 diffuses to the anode and is oxidized on the anode. The resulting current is proportional to the amount of generated H_2 .

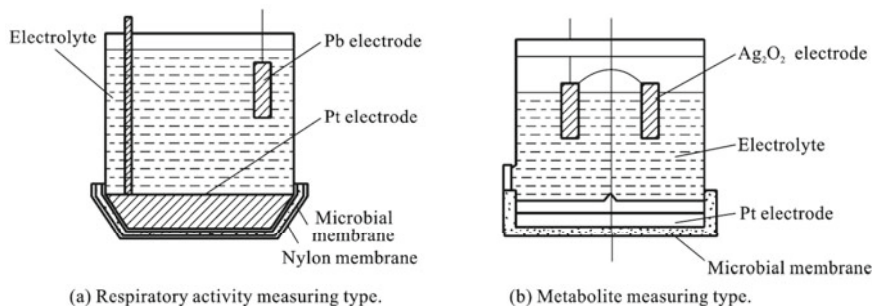


Fig. 7.102 Configurations of microorganism sensors

7.13.3 Immunosensor

Enzymatic and microorganism sensors mainly measure low-molecular organic compounds, and work badly for the polymer organic compounds. Using the recognition and binding functions of antibodies to antigens, immunosensors with high selectivity to proteins, polysaccharides and other polymers can be constructed. Immunosensors are based on immune response and can generally be divided into non-labeled immunosensors and labeled immunosensors.

1. Non-labeled immunosensors

Non-labeled immunosensors (also called direct immunoelectrodes) do not use any markers. The principle is that protein molecules (antigens or antibodies) carry a large amount of charge, when antigens and antibodies are combined, several electrochemical or electrical changes will occur. The parameters involved include: dielectric constant, electrical conductivity, membrane potential, ion permeability, ion concentration, etc. The occurrence of immune response can be detected by measuring the change of any of these parameters.

The non-labeled immunosensor is to form antigens and antibodies on the surface of the receptor, and convert the resulting physical changes into electrical signals. There are two types of sensors according to the measurement method: one is to fix the antibody (or antigen) on the surface of the membrane to become a receptor, and to measure the change of membrane potential before and after the immune response (Fig. 7.103a); the other is to fix the antibody (or antigen) to the surface of the metal electrode to become a receptor, and then measure the electrode potential change caused by the immune response (Fig. 7.103b).

The non-labeled immunosensor has the advantages of quick response and easiness to use. However, it has the disadvantages of requiring large amount of tested samples, low sensitivity and relatively high false positive rate due to the non-specific absorption.

2. Labeled immunosensors

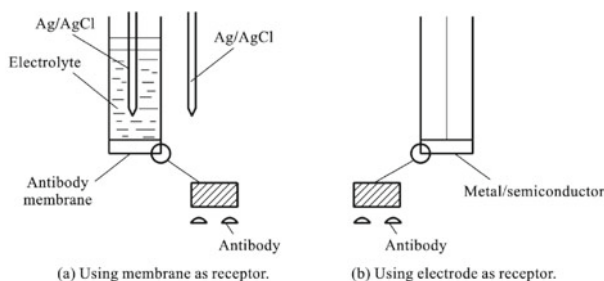


Fig. 7.103 Non-labeled immunosensors

Labeled immunosensors (also called indirect immunosensors) use enzymes, red blood cells, radioisotopes, stable free radicals, metals, liposomes, and salivary cells as markers. The principle is as follows: let a certain amount of labeled antigen and an equivalent amount of antibody to react, all the antigens will be combined with the antibody to form a complex; then take the same amount of labeled antigen and antibody as previously used, and then add the tested non-labeled antigen. At this time, because the labeled antigen and the non-labeled antigen compete with the antibody to form a complex, the amount of labeled antigen in the complex is changed (decreased or increased). Based on the change of labeled antigen, the original amount of non-labeled antigen can be inferred.

Compared with non-labeled immunosensors, labeled immunosensors have more applications. Some labeled immunosensors have been used in clinical analysis to determine concentrations of IgG and HCG, with a detection range up to 10^{-9} to 10^{-12} g mL⁻¹. This type of sensor requires a small amount of samples, generally only a few microliters to tens of microliters, with high sensitivity and good selectivity. However, it requires labeled antigen and the operation process is more complicate.

7.14 Selection of Sensors

How to select appropriate sensors for the practical applications is a very important question in the task of measurement. On the basis of the aforementioned measurement and sensor knowledge, the section criteria are briefly introduced.

1. Sensitivity

Generally speaking, the higher the sensitivity of the sensor, the better, because the higher the sensitivity, the smaller the amount of change can be perceived. Therefore, when a small change occurs in the measurement, the sensor has a larger output. However, it should be noted that the higher the sensitivity, the easier it is influenced by external interferences. It is necessary to consider not only detecting small values, but also preventing interference. Therefore, the sensitivity should be properly selected considering these two aspects.

2. Response characteristics

Within the measured frequency range, the response characteristics of the sensor must meet the conditions of non-distortion measurement. In addition, the response of the actual sensor always has a certain delay. In order to ensure that the measurement is not distorted, it is always hoped that the delay time is as short as possible.

Generally, physical type sensors that based on photoelectric effect and piezoelectric effect respond faster and have a wide operating frequency range. Structural type sensors, such as inductance, capacitance, and magnetoelectric sensors, are often limited by the inertia of the mechanical system in the structure. Their natural frequency is low, and the working frequency is low. In dynamic testing, the response

characteristics of the sensor have a direct impact on the test results. When selecting a sensor, the characteristics of the measured physical quantity (i.e. whether it is static or dynamic) should be fully considered.

3. Linear range

Any sensor has a certain linear range, and the output is proportional to the input within the linear range. The wider the linear range, the larger the working ranges of the sensor. To ensure the accuracy of measurement, sensors must work in the linear region. For example, for a force-measuring elastic system, the elastic limit of the material is the factor that determines the measurable force range. When the elastic limit is exceeded, an error would occur. At the same time, it should be noted that it is impossible for any sensor to have absolute linearity. The linearity only means it is approximately linear within an allowable limit. Therefore, when selecting a sensor, the variation range of the measured physical quantity must be considered, so that the linear error is within the allowable range.

4. Reliability

Reliability refers to the ability of instruments, devices and other products to perform specified functions within a specified time under specified conditions. Only the product performance parameters (especially the main performance parameters) are within the specified error range, can it be deemed to be able to complete the specified function.

In order to ensure high reliability of sensors in an application, a sensor with good design and manufacture process must be selected beforehand. During the usage, the specified use conditions should be strictly maintained to minimize the adverse effects. For example, for a resistance strain gauge, humidity will affect its insulation, temperature will affect its zero drift, and long-term use will cause creep. For variable-gap capacitive sensors, environmental humidity or oil immersed in the gap will change the dielectric constant of the medium. When the photosensitive surface of the photoelectric sensor has dust or moisture, it will change the luminous flux, polarization or spectral composition. For magnetoelectric sensors or Hall-effect elements, when working in an electric or magnetic field, errors would occur.

In mechanical engineering, some mechanical systems or automated processes often require sensors to be used for a long time without frequent replacement or calibration. The working environment is often harsh, with serious interference such as dust, oil, temperature, and vibration. For example, the gamma-ray detection device that controls the thickness of the steel plate in the hot rolling mill system, the force measurement system used in the adaptive grinding process or the automatic detection device of the part size. Under these circumstances, stricter requirements are put forward on the reliability of the sensor.

5. Accuracy

The accuracy of a sensor indicates the degree to which the output of the sensor is consistent with the true value. The sensor is at the input end of the test system.

Therefore, whether the sensor can truly reflect the measured value has a direct impact on the entire test system. However, the accuracy of the sensor is not chosen to be as high as possible, and price should also be considered. The higher the accuracy of a sensor, the more expensive it is. First, we should understand the purpose of the measurement and determine whether it is a qualitative analysis or a quantitative analysis. If it is a qualitative experimental study of relative comparison, only the relative comparison value needs to be obtained, and the absolute value is not required. If it is a quantitative analysis, an accurate value must be obtained, so the sensor is required to have a sufficiently high accuracy. For example, in order to study the positioning accuracy of moving parts of ultra-precision cutting machine tools, spindle rotation error, vibration and thermal deformation is often required to be measured with an accuracy within the range of 0.1–0.01 μm , so high-precision sensors must be used.

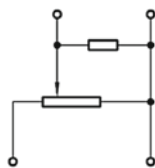
6. Measurement method

The working methods of the sensor, such as contact and non-contact testing, online and off-line testing are also important factors that should be considered when selecting a sensor. Different working methods have different requirements for sensors. In the testing of the moving parts of the mechanical system (such as the rotation error, vibration and torque of the rotating shaft), non-contact measurement is often required. Because the contact measurement would cause problems such as the wear of the measuring probe or the workpiece. The use of non-contact sensors such as capacitive sensor and eddy current sensor is more convenient. If resistance strain gauges are used, they need to be equipped with telemetry strain gauges or other devices.

Online testing is a testing method that is closer to the actual situation. In particular, the control and detection system of the automated process must be measured under real-time conditions on site. It is difficult to realize online detection, and has certain special requirements for sensors and test systems. For example, if you want to realize the online detection of surface roughness, the light section method, interferometric method and stylus-type contour detection method cannot be used. Instead, the laser detection method is preferable. The development of a new type of online detection sensor is also an important aspect of the current development of measurement technology.

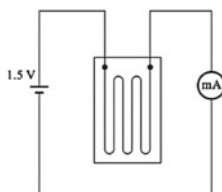
Exercises

- 7.1 What is the difference in working principle between resistance wire strain gauge and semiconductor strain gauge? What are the advantages and disadvantages of each? How should it be selected for specific applications?
- 7.2 Connect a rheostat sensor as shown in Exercise Fig. 7.1. What are the input and output? Under what conditions is there a good linear relationship between its output and input?



Exercise Fig. 7.1

- 7.3 There is a resistance strain gauge (Exercise Fig. 7.2), its sensitivity is $S = 2$, resistance is $R = 120 \Omega$, and the strain during working is $1000 \mu\epsilon$. What is the change of resistance $\Delta R = ?$ Suppose this strain gauge is connected to the circuit as shown in the figure, try to find: (1) the electric current when there is no strain; (2) the current when there is strain; (3) the relative change of the value of the ammeter; (4) try to analyze if the change can be read from the ammeter? (Note: $\mu\epsilon$ is micro strain)



Exercise Fig. 7.2

- 7.4 What are the factors influencing the sensitivity of inductive sensors (self-inductance type)? What measures can be taken to increase sensitivity? What are the consequences of taking these measures?
- 7.5 For a capacitance micrometer, the radius of sensor's circular plate is $r = 4 \text{ mm}$, the initial working gap $\delta_0 = 0.3 \text{ mm}$.
- (1) If the change of gap between the sensor and the workpiece is $\Delta\delta = \pm 1 \mu\text{m}$, what is the change of capacitance?
 - (2) If the sensitivity of the measuring circuit is $S_1 = 100 \text{ mV/pF}$ and the sensitivity of the reading instrument is $S_2 = 5 \text{ divisions/mV}$. When $\Delta\delta = \pm 1 \mu\text{m}$, how many divisions will be changed on the panel of the instrument?
- 7.6 What are the similarities and differences in the measurement circuits of capacitive sensors, inductive sensors, and resistive strain gauges?
- 7.7 Try to draw schematic diagrams to measure liquid pressure with capacitive, inductive, resistive and piezoelectric sensors. Then make comparisons between them.
- 7.8 What are the types of photoelectric sensors? What are the characteristics of each? What physical quantities can be measured with photoelectric sensors?

- 7.9 What is the Hall effect? Give three examples for the physical quantities that can be measured with Hall elements.
- 7.10 The sensitivity of a piezoelectric pressure sensor $S = 9 \times 105 \text{ pC/Pa}$ (pico-coulomb/Pa). Connect it to a charge amplifier whose sensitivity is adjusted to 0.005 V/pC , and the output of the amplifier is connected to an oscilloscope with sensitivity of 20 mm/V . Try to draw a block diagram of this measurement system, and calculate its total sensitivity.
- 7.11 Explain the differences in working principle of measuring pressure and displacement with optical fiber sensor.
- 7.12 What are the main types of thermal sensors? Briefly describe how they work.
- 7.13 What is the principle of thermocouple? When a measuring instrument is connected to the thermocouple circuit to measure the thermoelectric potential, will it affect the thermoelectric potential value of the original thermocouple circuit? Why?
- 7.14 What is the principle of thermistor? As far as you know, what parameters can be measured with thermistors? Try to describe the measurement principles for those physical quantities.
- 7.15 What are the main types of semiconductor gas sensors? Describe how they work.
- 7.16 What are the physical characteristics of ultrasound? Briefly describe the structure and working principle of the piezoelectric ultrasonic sensor. Give examples to illustrate the working principle of ultrasonic non-destructive testing.
- 7.17 There is a batch of turbine blades that needs inspection for cracks. Please list at least two methods to fulfill the task and explain the principles.
- 7.18 What are the two basic parts of a charge-coupled device (CCD)? Give an example to illustrate the measurement principle of the CCD sensor.
- 7.19 What is a biosensor? Briefly describe the basic structure and working principle of biosensors.
- 7.20 What are the basic principles for selecting sensors? How to apply these principles in practice? Give an example.

Chapter 8

Signal Conditioning Techniques



8.1 Overview of Signal Conditioning

Signal conditioning is the processing of analog signals to prepare it for the next stage of processing. Signal conditioning includes amplification, filtering, current to voltage conversion, voltage to current conversion, isolation, modulation and demodulation, etc.

In a digital measuring system, the analog signal of sensors should be converted into digital signal by an analog-to-digital converter (ADC), so that the result can be displayed in a computer or other digital devices. As shown in Fig. 8.1, signal conditioning techniques are implemented before the A/D conversion to prepare the analog signal for ADC, and their main purpose is to amplify the signal without distortion and increase the signal to noise ratio. Signal conditioning techniques decide the overall characteristics of the measuring system to a large extend.

Signal conditioning techniques are required before A/D conversion due to the following reasons.

(1) Weak sensor output

The output signal of some sensors is very weak, it cannot be distinguished by the ADC. For example, in structural health monitoring, piezoelectric transducer usually outputs a voltage in the magnitude of hundred microvolts or few millivolts. For a 10-bits ADC with full-scale input range of 5 V, the voltage resolution is 4.88 mV. A small fluctuation of hundred microvolts in the sensor output cannot be identified by the ADC. Thus, amplification is required before A/D conversion.

(2) Interference of noise

Nowadays, we are living in a space full of electronic devices and all kinds of electromagnetic waves. The electromagnetic field around a sensor will superimpose a noise to the original sensor output. Besides, the fluctuation of temperature and power supply voltage also causes noises in the sensor signal. Noise will obstruct our recognition of useful signal, especially when the sensor signal is very weak. Thus, filtering is required to improve the signal-to-noise ratio (SNR).

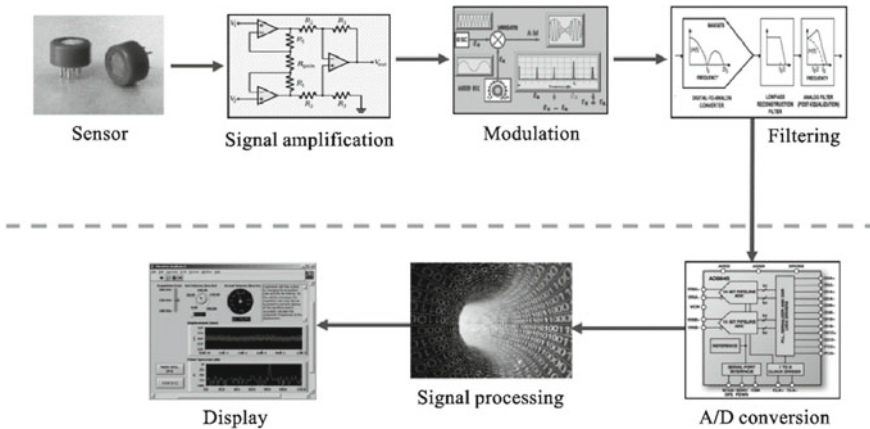


Fig. 8.1 Schematic diagram of a digital measuring system

(3) Long range transmission

Sometimes, sensor signals need to be transmitted in long range before they are analyzed and displayed. For example, in a factory, sensors are used to monitor the manufacturing process, but technicians are analyzing the sensor signals in a separate room. If the sensors are used for monitoring in the field, the signals need to be transmitted hundreds of kilometers to the city to be analyzed. During the long range transmission, signals are more easily to be distorted by noise. Thus, modulation and demodulation are used to increase the ability of long range transmission.

8.2 Analog Amplifiers and Operators

8.2.1 Operational Amplifier

Operational amplifier, sometimes abbreviated as op-amp or opamp, is an essential electronic component used in signal conditioning. It is a voltage amplifying device with very large gain. Its name comes from the previous usage in analog computers to implement mathematical operations.

An equivalent circuit of op-amp is shown in Fig. 8.2. It has two input terminals, one inverting input terminal which is marked with a minus sign and one non-inverting input terminal which is marked with a plus sign. The name “inverting” and “non-inverting” comes from the phase shift between output and input signals. When the inverting input terminal is fed by a sinusoidal signal while the non-inverting input terminal is grounded, the output signal is inverted, i.e. it has a phase shift of 180° . If we fed the same signal to the non-inverting input terminal and ground the inverting input terminal, the output signal will have the same phase as input.

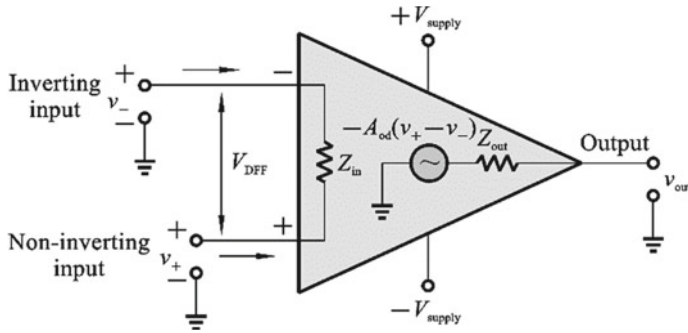


Fig. 8.2 Equivalent circuit of op-amp

The output voltage at the output terminal is an amplification of the voltage difference between two input terminals:

$$v_{\text{out}} = A_{\text{od}} \cdot (v_+ - v_-) \quad (8.1)$$

where v_{out} is the output voltage, v_- and v_+ are the voltages in inverting and non-inverting input terminals respectively, A_{od} is the open loop gain.

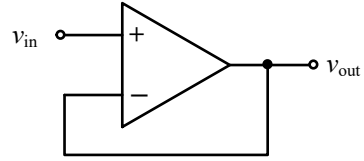
Virtual short and virtual open are the most important concepts for analyzing op-amp circuit. These concepts originate from the characteristics of op-amp. An ideal op-amp has infinite open loop gain (typical values range from 65 to 100 dB for practical op-amp) and infinite input impedance (typical values range from 0.5 to 2 M Ω for practical op-amp). It can be noticed from Eq. (8.1) that, if the output voltage is a finite value while the open loop gain is infinite, then the term $v_+ - v_-$ must be zero. This is equivalent to short the two input terminals, hence this phenomenon is called virtual short. Because of the huge input impedance, no current will flow into the op-amp (practical op-amps have small leakage current in the magnitude of few nanoamperes to few microamperes). This is equivalent to open the op-amp and input circuits, hence this phenomenon is called virtual open.

8.2.2 Typical Amplification and Operation Circuits

1. Voltage follower

Voltage follower, which is also called isolator or unity-gain buffer, is a circuit with its output voltage equals the input voltage. The most basic voltage follower only consists of one op-amp as shown in Fig. 8.3. The output terminal is directly connected to the inverting input terminal to form a negative feedback. Without the negative feedback, a small voltage in the input would cause a huge output voltage. While with the negative feedback, compensation is made by returning the increased output voltage to the differential input voltage as a negative value. Finally, equilibrium is reached

Fig. 8.3 Voltage follower circuit



when the output voltage equals the input voltage. This circuit can also be analyzed using the concept of virtual short. Since the output terminal is directly connected to the inverting input terminal, the output voltage equals the voltage at the inverting input terminal ($v_{out} = v_-$). And due to virtual short, the voltage at the inverting input terminal equals the input voltage at the non-inverting input terminal ($v_{in} = v_-$). As a result, the output voltage equals the input voltage.

The voltage follower neither increases the voltage amplitude nor decreases the noise in the signal, but it plays particular roles. The voltage that has been transmitted from a sensor or a previous sub-circuit to the subsequent sub-circuit depends on the output impedance of the sensor or previous sub-circuit (Z_{out-1}) and the input impedance of the subsequent sub-circuit (Z_{in-2}). The larger the ratio Z_{in-2}/Z_{out-1} , the larger the voltage is transmitted. The voltage follower has very large input impedance, which makes it suitable to extract the output signal of a sensor. The large input impedance also makes it a good isolator to reduce the interferences between sensor and measuring instrument. Besides, the relatively small output impedance increases the ability to drive an actuator.

2. Inverting amplifier

Inverting amplifier is one of the most commonly used circuits for amplifying signals. As shown in Fig. 8.4, the non-inverting input terminal is grounded, thus $v_+ = 0$. Due to virtual short, the voltage at the inverting input terminal is also zero, $v_- = 0$. Due to virtual open, there is no current flowing into the inverting input terminal. Thus, the current flowing through R_1 is equal to the current flowing through R_f . Then, the following expressions can be obtained by Ohm's law:

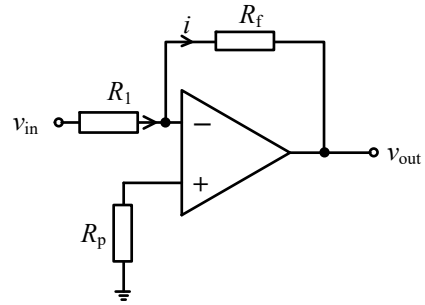
$$v_{in} = i \cdot R_1 + v_- = i \cdot R_1 \quad (8.2)$$

$$v_{out} = v_- - i \cdot R_f = -i \cdot R_f \quad (8.3)$$

Combining Eqs. (8.2) and (8.3), the relationship between input and output voltages can be found:

$$v_{out} = -\frac{R_f}{R_1} \cdot v_{in} \quad (8.4)$$

The voltage gain of the circuit is:

Fig. 8.4 Inverting amplifier

$$A_v = -\frac{R_f}{R_1} \quad (8.5)$$

If the feedback resistance R_f is larger than the input resistance R_1 , the amplitude of output signal will be larger than that of the input signal. We can control the gain by replacing the feedback resistor.

It can be noticed from Eq. (8.4) that, besides amplitude amplification, the circuit also inverts the signal and introduces a phase shift of 180° to the sinusoidal signals. This is why this circuit is called inverting amplifier and it is due to the connection of input signal to the inverting input terminal.

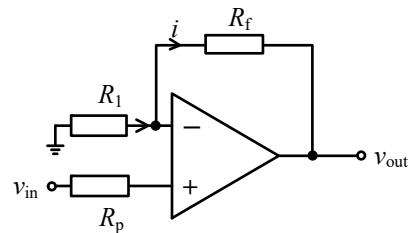
3. Non-inverting amplifier

The phase shift caused by the amplifier is sometimes undesirable. It can be avoided by feeding the input signal to the non-inverting input terminal of the op-amp. The non-inverting amplifier is shown in Fig. 8.5. The currents flowing through R_1 and R_f are respectively:

$$i_1 = -\frac{v_-}{R_1} \quad (8.6)$$

$$i_f = \frac{v_- - v_{out}}{R_f} \quad (8.7)$$

Due to virtual open, there is no current flowing into the non-inverting input terminal. Thus, Eq. (8.6) equals Eq. (8.7), and the following relationship can be

Fig. 8.5 Non-inverting amplifier

obtained:

$$v_{\text{out}} = \left(1 + \frac{R_f}{R_1}\right)v_- \quad (8.8)$$

Then, due to virtual short, we get:

$$v_- = v_+ = v_{\text{in}} \quad (8.9)$$

Finally, the relationship between input and output voltages is given as:

$$v_{\text{out}} = \left(1 + \frac{R_f}{R_1}\right)v_{\text{in}} \quad (8.10)$$

It can be noticed from Eq. (8.10) that the output voltage is in phase with the input voltage. However, there is a disadvantage of the non-inverting amplifier, i.e. the output signal is not proportional to the feedback resistance.

4. Adder/subtractor

Adder and subtractor are analog circuits that implement mathematical operations of addition and subtraction to multi-channel input signals. A basic adder is shown in Fig. 8.6. Due to virtual open, the current through R_f is:

$$i_f = \sum_{k=1}^N i_k$$

According to Ohm's law, we get:

$$\frac{v_1}{R_1} + \frac{v_2}{R_2} + \dots + \frac{v_N}{R_N} = -\frac{v_{\text{out}}}{R_f}$$

Thus, the output voltage can be expressed as:

Fig. 8.6 An adder circuit

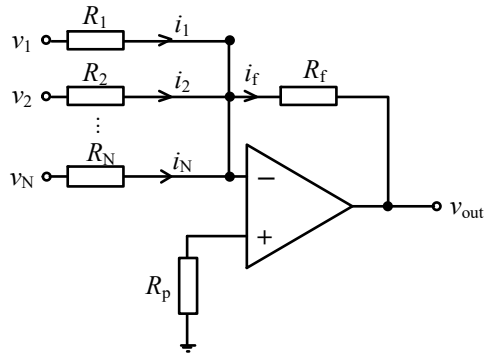
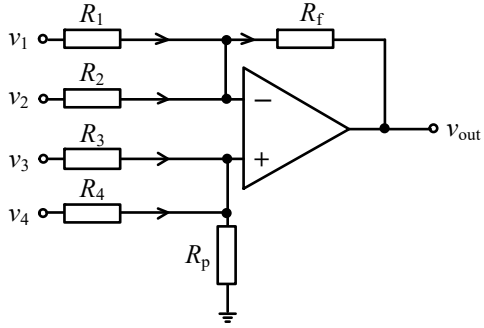


Fig. 8.7 A circuit for mixed operations of addition and subtraction



$$v_{\text{out}} = -R_f \left(\frac{v_1}{R_1} + \frac{v_2}{R_2} + \cdots + \frac{v_N}{R_N} \right) \quad (8.11)$$

The output voltage is the addition of weighted input voltages.

A circuit for mixed operations of adding and subtracting can be designed by feeding some input signals to the non-inverting input terminal and others to the inverting input terminal. An example with four input signals is shown in Fig. 8.7. The relationship between output voltage and input voltages can be obtained by using the principles of virtual short and virtual open:

$$v_{\text{out}} = R_f \left(\frac{v_3}{R_3} + \frac{v_4}{R_4} - \frac{v_1}{R_1} - \frac{v_2}{R_2} \right) \quad (8.12)$$

5. Integrator

An integrator (also called integral amplifier) is a circuit that integrates the input signal (Fig. 8.8). It can be used as a waveform convertor, e.g. inputting a square wave to the integrator will get a triangular wave. It can also be used as a phase shifter to integral a sine signal into a cosine signal to change its phase. A basic integrator is shown in.

The current through the capacitor equals the current through the resistor due to virtual open:

$$i_C = i_R = \frac{v_{\text{in}}}{R} \quad (8.13)$$

The output voltage is the integration of current through the capacitor:

$$v_{\text{out}} = -\frac{1}{C} \int i_C dt \quad (8.14)$$

Substituting Eq. (8.13) into Eq. (8.14), we get:

$$v_{\text{out}} = -\frac{1}{RC} \int v_{\text{in}} dt \quad (8.15)$$

Within the time interval between t_0 to t_1 , the output voltage is

$$v_{\text{out}}(t_1) = -\frac{1}{RC} \int_{t_0}^{t_1} v_{\text{in}}(t) dt + v_{\text{out}}(t_0) \quad (8.16)$$

where $v_{\text{out}}(t_1)$ is the output voltage at time t_1 , $v_{\text{in}}(t)$ is the time-varying input voltage, $v_{\text{out}}(t_0)$ is the output voltage at time t_0 . It can be noticed from Eq. (8.16) that the output voltage depends not only on the current input but also on the historical inputs.

6. Differentiator

A differentiator has the converse effect of an integrator. A basic differentiator circuit is obtained by switching the resistor and the capacitor in Fig. 8.9. The current through the capacitor equals the current through the resistor due to virtual open:

$$i_R = i_C = C \frac{dv_{\text{in}}}{dt}$$

Hence, the output voltage is:

$$v_{\text{out}} = -i_R R = -RC \frac{dv_{\text{in}}}{dt} \quad (8.17)$$

Fig. 8.8 An integral amplifier

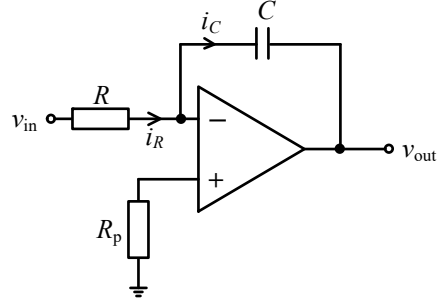
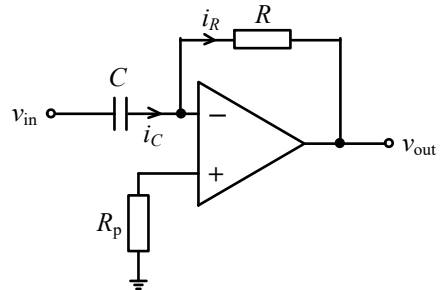


Fig. 8.9 A differentiator



7. Logarithmic amplifier

A logarithmic amplifier outputs a voltage which is proportional to the logarithmic of the input voltage. It can be designed with a diode due to its exponential voltage-current characteristic under forward bias:

$$i = I_S(e^{\frac{u}{u_T}} - 1) \quad (8.18)$$

where i is the diode current, u is the voltage across diode, I_S is the saturation current, e is the Euler's number which approximately equals 2.7128, and u_T is

$$u_T = \frac{q}{kT}$$

where q is the elementary charge, k is the Boltzmann constant, T is temperature in Kelvin. At room temperature, $u_T \approx 0.26$ mV. When the voltage across diode is much larger than u_T , Eq. (8.18) can be approximated as:

$$i \approx I_S e^{\frac{u}{u_T}} \quad (8.19)$$

A logarithmic amplifier with diode is shown in Fig. 8.10. According to Eq. (8.19), the voltage across the diode is:

$$u_D = u_T \ln \frac{i_D}{I_S} \quad (8.20)$$

Due to virtual short, $v_- = v_+ = 0$. Then, due to virtual open,

$$i_D = i_R = \frac{v_{in}}{R} \quad (8.21)$$

By combining Eqs. (8.20) and (8.21), the output voltage is:

$$v_{out} = -u_D = -u_T \ln \frac{v_{in}}{I_S R} \quad (8.22)$$

Fig. 8.10 A logarithmic amplifier based on diode

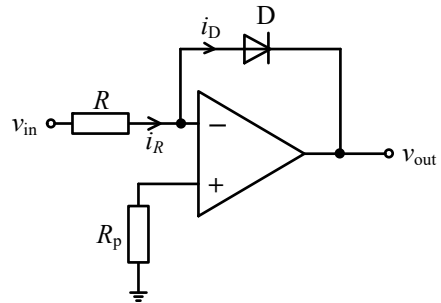
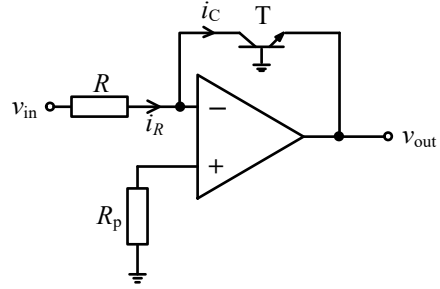


Fig. 8.11 A logarithmic amplifier based on transistor



A transistor can also be used to build a logarithmic amplifier as shown in Fig. 8.11. When the voltage across base-emitter junction v_{BE} is much larger than u_T , and the gain α is approximately equal to 1, the collector current and emitter current satisfy:

$$i_C = \alpha i_E \approx I_S e^{\frac{v_{BE}}{u_T}}$$

Thus,

$$v_{BE} \approx u_T \ln \frac{i_C}{I_S}$$

The output voltage for the circuit in Fig. 8.11 is:

$$v_{out} = -v_{BE} = -u_T \ln \frac{v_{in}}{I_S R} \quad (8.23)$$

It can be found from Eqs. (8.22) and (8.23) that the output voltages of the circuits in Figs. 8.10 and 8.11 are the same.

8. Exponential amplifier

An exponential amplifier, as shown in Fig. 8.12, has the converse effect of a logarithmic amplifier. It can be obtained by switching the positions of the transistor and the resistor in Fig. 8.11. Due to virtual short, $v_- = v_+ = 0$, thus the voltage across base-emitter junction is

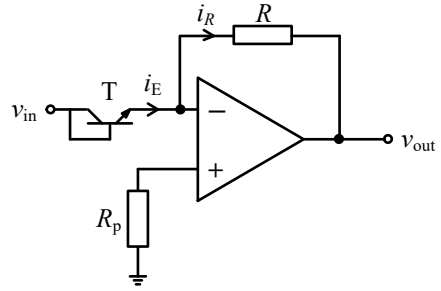
$$v_{BE} = v_{in}$$

Due to virtual open,

$$i_R = i_E \approx I_S e^{\frac{v_{in}}{u_T}}$$

Thus, the output voltage is

$$v_{out} = -i_R R = -I_S R e^{\frac{v_{in}}{u_T}} \quad (8.24)$$

Fig. 8.12 An exponential amplifier

9. Multiplier

A multiplier outputs the multiplication of two input signals. The operation of multiplication is converted into the logarithmic, exponential and addition operations as:

$$v_1 \times v_2 = e^{(\ln v_1 + \ln v_2)} \quad (8.25)$$

The analog multiplier circuit is designed according to Eq. (8.25) and it is shown in Fig. 8.13.

According to Eq. (8.23), the output voltages of the logarithmic amplifiers are:

$$v_{o1} = -u_T \ln \frac{v_1}{I_S R}$$

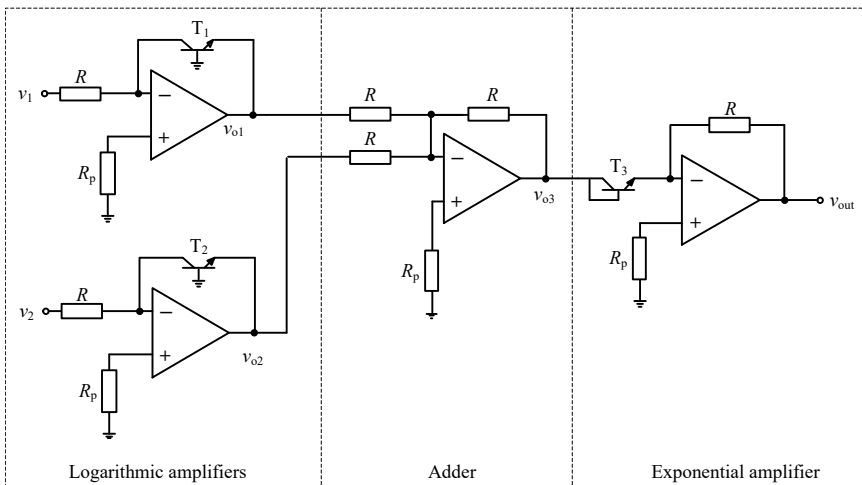
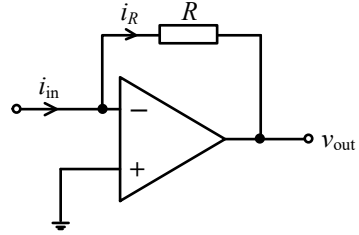
**Fig. 8.13** A multiplier circuit

Fig. 8.14 A current to voltage converter



$$v_{o2} = -u_T \ln \frac{v_2}{I_S R}$$

According to Eq. (8.11), the output voltage of the adder is:

$$v_{o3} = u_T \ln \frac{v_1 v_2}{(I_S R)^2}$$

According to Eq. (8.24), the final output voltage of the multiplier is:

$$v_{out} = -\frac{v_1 v_2}{I_S R} \quad (8.26)$$

If the adder in Fig. 8.13 is replaced by a subtractor, then the circuit becomes a divider to calculate the ratio of two input voltages.

10. Current to voltage convertor and voltage to current convertor

A basic current to voltage convertor is shown in Fig. 8.14. According to Ohm's law, the output voltage is:

$$v_{out} = i_R R + v_-$$

Due to virtual short, we get $v_- = 0$ and due to virtual open, we get $i_R = i_{in}$. Finally, the output voltage relates with the input current by:

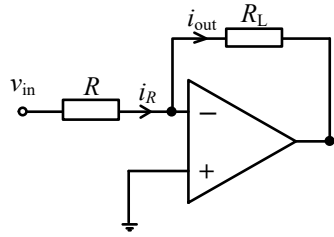
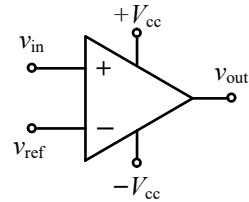
$$v_{out} = i_{in} R \quad (8.27)$$

A basic voltage to current convertor is shown in Fig. 8.15. The output current is the current through the load resistance. Due to virtual open,

$$i_{out} = i_R = \frac{v_{in}}{R} \quad (8.28)$$

11. Voltage comparator

A voltage comparator compares the input voltage with a reference voltage and outputs a digitalized (binarized) signal. A basic voltage comparator is shown in Fig. 8.16.

Fig. 8.15 A voltage to current convertor**Fig. 8.16** A voltage comparator

Without negative feedback, the output voltage of the op-amp is

$$v_{out} = G_{inf}(v_{in} - v_{ref})$$

where G_{inf} is the infinite gain of the op-amp. Thus the op-amp is supposed to output a positive infinity when input voltage is larger than the reference voltage and a negative infinity when input voltage is smaller than the reference voltage. However, due to the finite voltage supply, the output voltage of a practical op-amp is

$$v_{out} = V_{CC} \cdot \text{sign}(v_{in} - v_{ref}) \quad (8.29)$$

An example of input and output voltages of a voltage comparator is shown in Fig. 8.17. In this example, the reference is set to zero by grounding the inverting input terminal.

12. Charge amplifier

Charge amplifier is a circuit that outputs a voltage proportional to the integrated input charge or current. The charge amplifier has a capacitor in the feedback loop to implement the integration, which is similar to an integrator. But, the input signal to the charge amplifier is charge or current. It is mainly used for piezoelectric sensors in engineering measurement. A basic charge amplifier for piezoelectric transducer

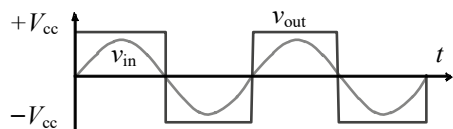
Fig. 8.17 An example of input and output voltages of a voltage comparator

Fig. 8.18 A charge amplifier for piezoelectric transducer

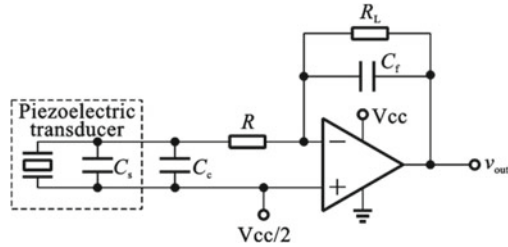
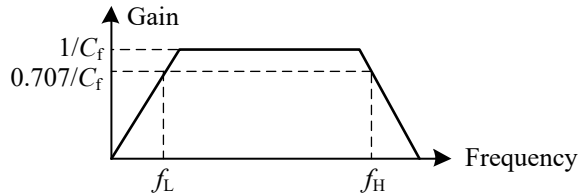


Fig. 8.19 Frequency response of the charge amplifier



is shown in Fig. 8.18. The transducer is equivalent to a charge source connected in parallel with its capacitance, and C_c is the capacitance introduced by cable.

Due to the existence of capacitances, the circuit only works for signals within a bandwidth as shown in Fig. 8.19. The cutoff frequencies are respectively:

$$f_L = \frac{1}{2\pi R_f C_f}$$

$$f_H = \frac{1}{2\pi R(C_s + C_c)}$$

8.3 Bridge Circuits

A bridge is a measuring circuit that converts changes in resistance, inductance, and capacitance into changes of voltage or current. It is also known as Wheatstone bridge. According to the position of sensor in the bridge circuit, bridges can be divided into quarter bridge, half bridge and full bridge, as shown in Fig. 8.20. They differ each other by the number of sensors in the bridge and the positions.

According to the type of the excitation voltage, bridges can be divided into DC bridge and AC bridge. A bridge powered by a DC power supply is called a DC bridge, and a bridge powered by an AC power supply is called an AC bridge.

According to the output mode, bridges can be divided into balanced bridge and unbalanced bridge. Bridges that meet Wheatstone balance condition are balanced bridges, otherwise are unbalanced bridges.

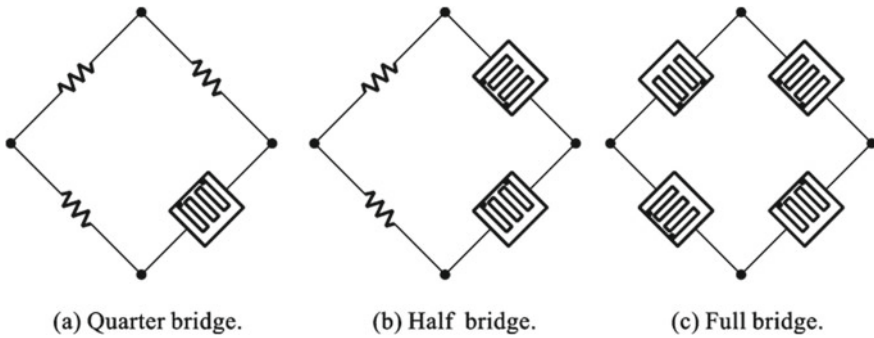


Fig. 8.20 Different types of bridge circuits

8.3.1 DC Bridge

Figure 8.21 shows the basic form of a DC bridge. In the figure, R_1, R_2, R_3, R_4 are called the bridge arm resistance, and e_0 is the DC voltage supply for the bridge.

When the bridge output terminals b and d are connected to an instrument or amplifier with relatively large input impedance, it can be regarded as an open circuit, the output current is zero, and the output voltage is e_y . The currents in the circuit are:

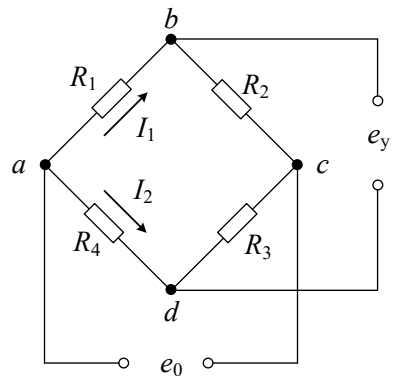
$$I_1 = \frac{e_0}{R_1 + R_2} \quad (8.30)$$

$$I_2 = \frac{e_0}{R_3 + R_4} \quad (8.31)$$

The potential drops between a, b and a, d are respectively:

$$U_{ab} = I_1 R_1 = \frac{R_1}{R_1 + R_2} e_0 \quad (8.32)$$

Fig. 8.21 DC bridge



$$U_{ad} = I_2 R_4 = \frac{R_4}{R_3 + R_4} e_0 \quad (8.33)$$

Thus the output voltage is:

$$e_y = U_{ab} - U_{ad} = \left(\frac{R_1}{R_1 + R_2} - \frac{R_4}{R_3 + R_4} \right) e_0 = \frac{R_1 R_3 - R_2 R_4}{(R_1 + R_2)(R_3 + R_4)} e_0 \quad (8.34)$$

When the output is zero, the following condition should be satisfied:

$$R_1 R_3 = R_2 R_4 \quad (8.35)$$

Equation (8.35) is called the balance condition of the DC bridge. Appropriate selection of the resistance of each bridge arm can make the bridge meet the balance condition before measurement, i.e. the output voltage $e_y = 0$.

If the resistance of R_1 changes by ΔR , then the output voltage is:

$$e_y = \left(\frac{R_1 + \Delta R}{R_1 + \Delta R + R_2} - \frac{R_4}{R_3 + R_4} \right) e_0 \quad (8.36)$$

In practical bridges, usually the resistances of four arms are equal in the initial state:

$$R_1 = R_2 = R_3 = R_4 = R \quad (8.37)$$

Then the output becomes:

$$e_y = \frac{\Delta R}{4R + 2\Delta R} e_0 \quad (8.38)$$

Usually the change of resistance is much smaller than its original value, i.e. $\Delta R \ll R$, thus the term $2\Delta R$ in the denominator can be ignored. And we get:

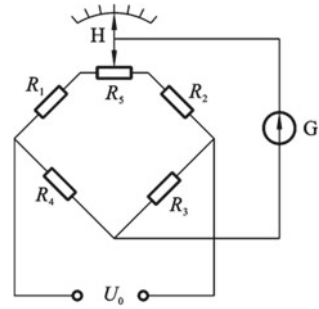
$$e_y = \frac{\Delta R}{4R} e_0 \quad (8.39)$$

It can be seen that the output voltage is proportional to the voltage of power supply. Under the condition of $\Delta R \ll R$, the output voltage is proportional to the relative change of resistance $\Delta R/R$. The sensitivities of the three types of bridges shown in Fig. 8.20 are:

$$S_1 = \frac{e_y}{\Delta R/R} = \frac{1}{4} e_0 \quad (8.40)$$

$$S_2 = \frac{e_y}{\Delta R/R} = \frac{1}{2} e_0 \quad (8.41)$$

Fig. 8.22 Null measurement method of bridge



$$S_3 = \frac{e_y}{\Delta R / R} = e_0 \quad (8.42)$$

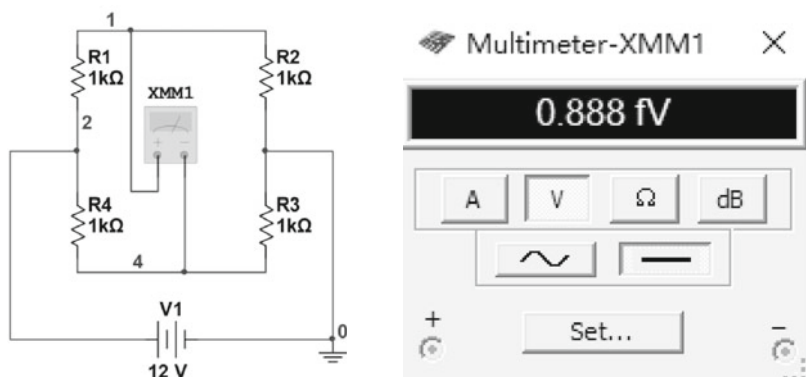
where S_1 , S_2 and S_3 are sensitivities for quarter bridge, half bridge and full bridge respectively.

The null measurement method for bridge is shown in Fig. 8.22. Before the measurement, the bridge is adjusted to balance, thus the reading on the ammeter G is zero. When one of the arms is used for measurement, the resistance of corresponding arm changes. And the ammeter G shows a non-zero value. Adjust the potentiometer H to make the reading of ammeter G go back to zero again. The scale on the potentiometer H is proportional to the change in the resistance of the bridge arm, so the indicated value of H can directly express the measured value. In this method, we are adjusting the bridge to make the reading on G to be zero, thus it is called null measurement.

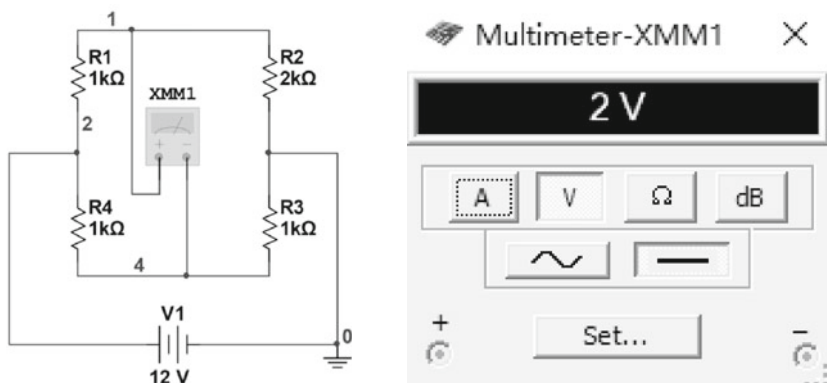
Example 8.1 Multisim Simulation for DC Bridge A DC bridge is built in Multisim as shown in Fig. 8.23a. When the resistances of four arms are all equal to 1 k Ω , the bridge is under balance, and the output voltage is shown in Fig. 8.23b. Thus, the output voltage is approximately 0 V when the bridge is under balance. The small reading in shown in the software is due to numerical errors. When resistance of one arm is changed to 2 k Ω , the reading shown on the voltmeter is 2 V, which is the same as the theoretical calculation.

8.3.2 AC Bridge

The AC bridge circuit is shown in Fig. 8.24. Its excitation e_0 is AC voltage. The four bridge arms can be combination of inductance L , capacitance C or resistance R , and is represented by impedance Z . If impedance, current and voltage are all represented by complex numbers, the balance relation of DC bridge is also applicable in AC bridge, namely the balance of AC bridge must satisfy



(a)



(b)

Fig. 8.23 Multisim simulation for DC bridge

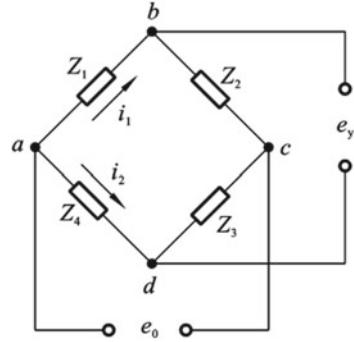
$$Z_1 Z_3 = Z_2 Z_4 \quad (8.43)$$

The complex impedance contains amplitude and phase information, and the impedances can be expressed in exponential form as

$$\begin{aligned} Z_1 &= Z_{01} e^{j\varphi_1} & Z_2 &= Z_{02} e^{j\varphi_2} \\ Z_3 &= Z_{03} e^{j\varphi_3} & Z_4 &= Z_{04} e^{j\varphi_4} \end{aligned} \quad (8.44)$$

Then the balance condition for AC bridge becomes:

$$Z_{01} Z_{03} e^{j(\varphi_1 + \varphi_3)} = Z_{02} Z_{04} e^{j(\varphi_2 + \varphi_4)} \quad (8.45)$$

Fig. 8.24 AC bridge

where Z_{01} , Z_{02} , Z_{03} and Z_{04} are modules of impedance, φ_1 , φ_2 , φ_3 and φ_4 are the phases between voltage and current. When pure resistors are used, the phases are zero. When inductive impedance is used, voltage is ahead of current, and $\varphi > 0$; when capacitive impedance is used, voltage is behind current, and $\varphi < 0$.

According to the calculations of DC bridge, we can get:

$$U_y = \frac{Z_1 Z_3 - Z_2 Z_4}{(Z_1 + Z_2)(Z_3 + Z_4)} U_0 \quad (8.46)$$

Under balance condition:

$$\begin{cases} Z_{01} Z_{03} = Z_{02} Z_{04} \\ \varphi_1 + \varphi_3 = \varphi_2 + \varphi_4 \end{cases} \quad (8.47)$$

The output voltage is zero.

1. Capacitive AC bridge

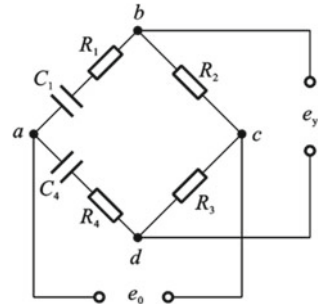
A capacitive AC bridge is shown in Fig. 8.25. Two of the arms are serial connections of resistor and capacitor. The balance condition for this bridge is:

$$\begin{aligned} \left(R_1 + \frac{1}{j\omega C_1} \right) R_3 &= \left(R_4 + \frac{1}{j\omega C_4} \right) R_2 \\ R_1 R_3 + \frac{R_3}{j\omega C_1} &= R_2 R_4 + \frac{R_2}{j\omega C_4} \end{aligned} \quad (8.48)$$

Let the real and imaginary components to be equal respectively, we get:

$$\begin{aligned} R_1 R_3 &= R_2 R_4 \\ \frac{R_3}{C_1} &= \frac{R_2}{C_4} \end{aligned} \quad (8.49)$$

Fig. 8.25 Capacitive AC bridge



Example 8.2 Multisim Simulation for Capacitive AC Bridge A capacitive AC bridge is built in Multisim as shown in Fig. 8.26. The excitation is a sinusoid with amplitude of 10 V and frequency of 100 Hz. An oscilloscope is added in the circuit to observe the output voltage. When the bridge is unbalanced, a sinusoidal signal can be observed.

2. Inductive AC bridge

An inductive AC bridge is shown in Fig. 8.27. Two of the arms are serial connections of resistor and inductor. The balance condition for this bridge is:

$$\begin{aligned} R_1 R_3 &= R_2 R_4 \\ L_1 R_3 &= L_4 R_2 \end{aligned} \quad (8.50)$$

Some conclusions can be inferred for the AC bridges:

- (1) If two adjacent bridge arms are resistors, according to the balance condition, the other two bridge arms must have the same type of impedance. Either both are capacitive or both are inductive;
- (2) If two opposite arms of the bridge are resistors, according to the balance condition, the other two bridge arms must have different types of impedance. If one arm is capacitive, the opposite side should be inductive.

If four arms are all resistors, then $\varphi_1 = \varphi_2 = \varphi_3 = \varphi_4 = 0$. If other factors are ignored, then the balance condition of AC bridge is the same as DC bridge. However, in the actual use of the AC bridges, there are more factors that affect errors of AC bridge than errors of DC bridge. Since the balance of the AC bridge must meet both the amplitude and impedance conditions, the adjustment of AC bridge balance is more complicate than that of DC bridge.

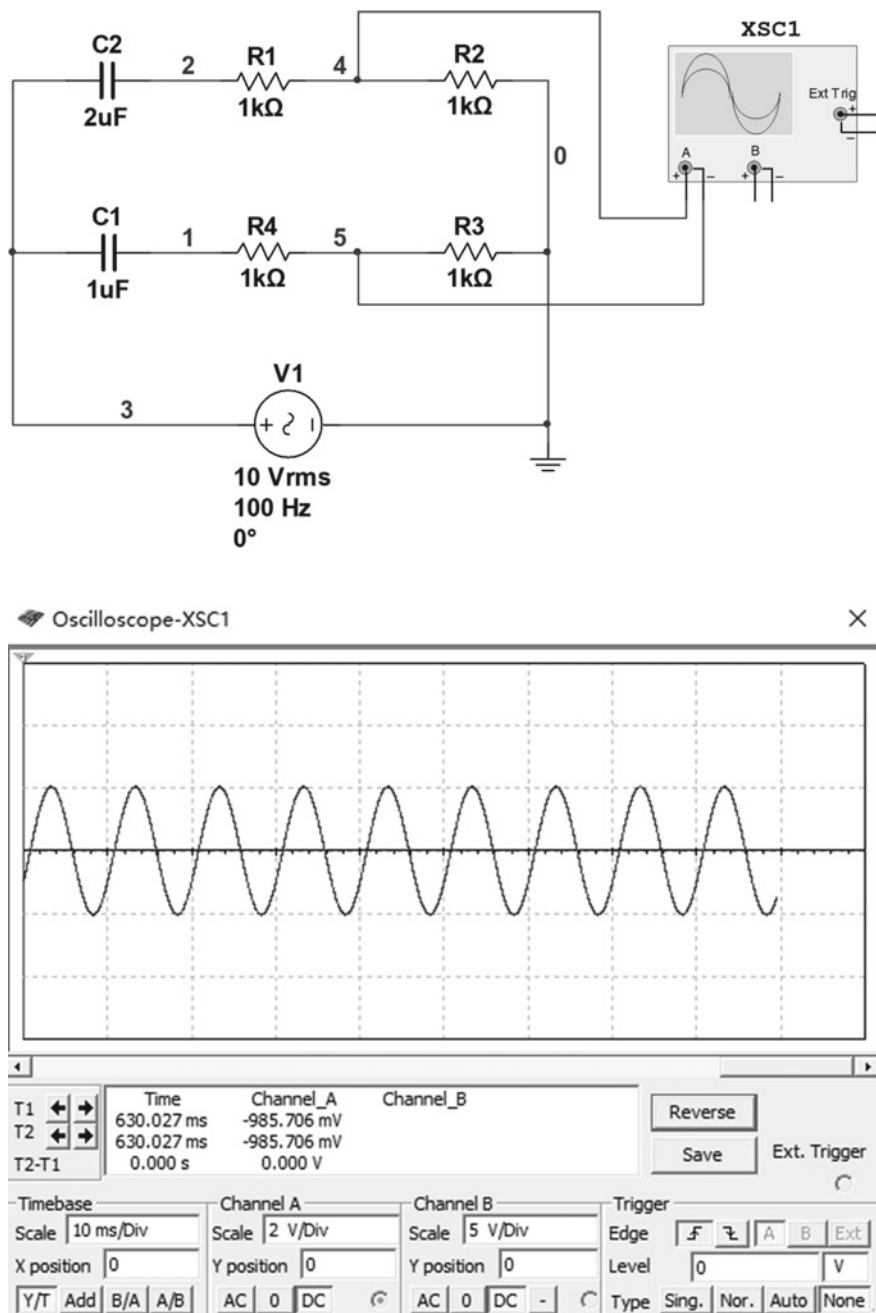
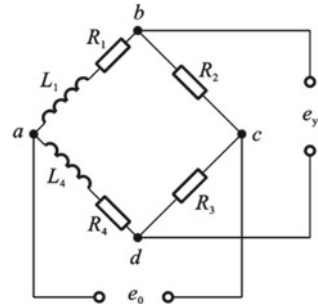


Fig. 8.26 Multisim simulation for capacitive AC bridge

Fig. 8.27 Inductive AC bridge



8.4 Analog Filters

Noise always superimpose with the sensor signal to cause interferences. Filters are used to remove or reduce noise. Filters can be classified into analog filters and digital filters. Digital filters are implemented by CPU, DSP, MCU, etc. after the analog-to-digital conversion, and they have been introduced in Chap. 6. In this section, analog filters are introduced. They are implemented by op-amp, resistors and capacitors before the analog-to-digital conversion. Analog filters can be further classified as low-pass filter, high-pass filter, band-pass filter and band-stop filter according to the frequency components that have been attenuated.

8.4.1 Low-Pass Filter

Low-pass filter is an electronic circuit that allows the pass of low-frequency signals and attenuates signals with frequency higher than the cutoff frequency. A simple low-pass filter can be constructed with only a resistor and a capacitor as shown in Fig. 8.28. This circuit does not require a power supply, hence it is called a passive filter.

Assuming the input and output of the filter are v_{in} and v_{out} respectively, the governing equation of the circuit is:

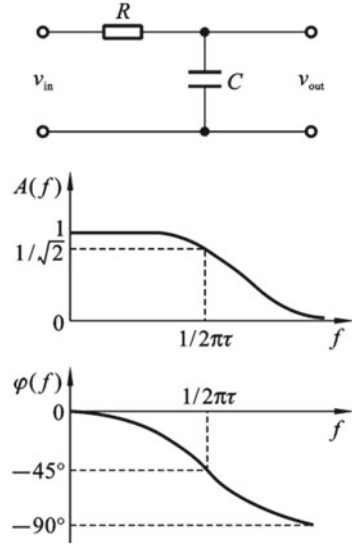
$$RC \frac{dv_{out}}{dt} + v_{out} = v_{in} \quad (8.51)$$

Under sinusoidal input, the current in the circuit is:

$$i = \frac{v_{in}}{Z_R + Z_C} = \frac{v_{in}}{R - j \frac{1}{\omega C}}$$

The voltage across the capacitor, i.e. the output voltage, is:

Fig. 8.28 A typical passive low-pass filter



$$v_{out} = i \cdot Z_C = \frac{-j \frac{1}{\omega C}}{R - j \frac{1}{\omega C}} v_{in}$$

The transfer function is:

$$H(j\omega) = \frac{v_{out}}{v_{in}} = \frac{-j \frac{1}{\omega C}}{R - j \frac{1}{\omega C}} = \frac{1}{1 + j\omega RC} \quad (8.52)$$

Then, the magnitude and phase of the transfer function can be derived from Eq. (8.52):

$$A(j\omega) = |H(j\omega)| = \frac{1}{\sqrt{1 + (\omega RC)^2}} \quad (8.53)$$

$$\varphi(j\omega) = \arctan \frac{\text{Im}(H(j\omega))}{\text{Re}(H(j\omega))} = -\arctan(\omega RC) \quad (8.54)$$

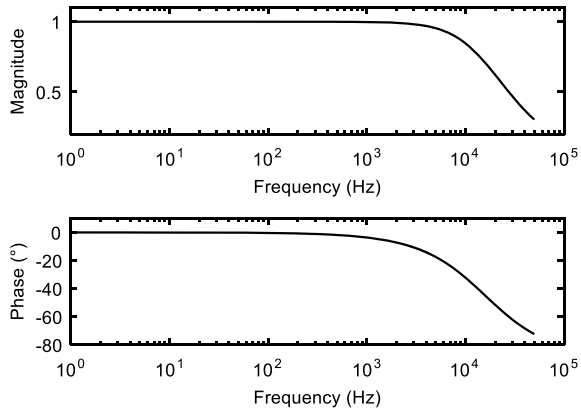
or

$$A(f) = |H(f)| = \frac{1}{\sqrt{1 + (2\pi f RC)^2}} \quad (8.55)$$

$$\varphi(f) = -\arctan(2\pi f RC) \quad (8.56)$$

The corresponding magnitude and phase spectra can be plotted using Eqs. (8.55) and (8.56). An example is shown in Fig. 8.29 with $R = 10 \, \Omega$ and $C = 1 \, \mu\text{F}$. Notice

Fig. 8.29 The magnitude and phase spectra of a low-pass filter



that the horizontal axis in Fig. 8.29 is the frequency, $f = \omega/2\pi$, and it is plotted in logarithmic scale.

The time constant for the R–C circuit in Fig. 8.28 is $\tau = RC$. The cutoff angular frequency of the low-pass filter is defined as:

$$\omega_L = \frac{1}{\tau} = \frac{1}{RC}$$

And the corresponding cutoff frequency is:

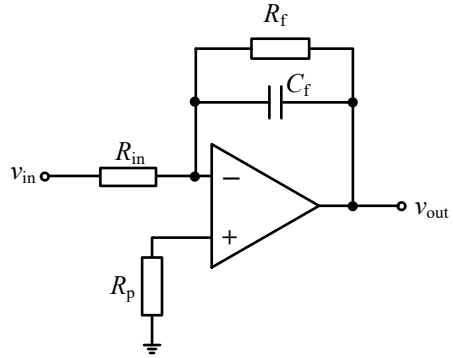
$$f_L = \frac{\omega_L}{2\pi} = \frac{1}{2\pi RC} \quad (8.57)$$

At the cutoff frequency, the magnitude is $A(j\omega_L) = 1/\sqrt{2} = 0.707$, which is approximately equivalent to a reduction of 3 dB in the magnitude. For the resistance and capacitance values given in Fig. 8.29, the cutoff frequency is $f_L = 15.9$ kHz. When $f \ll 1/(2\pi\tau)$, $A(f) = 1$ and $\varphi = 0$. Under this condition, signals can pass without attenuation, and the RC filter can be considered as an undistorted system. When $f = 1/(2\pi\tau)$, $A(f) = 0.707$, it corresponds to the point of -3 dB decay. When $f \gg 1/(2\pi\tau)$, the output is proportional to the integral of input:

$$v_{\text{out}} = \frac{1}{RC} \int v_{\text{in}} dx$$

At this time, the RC filter acts as an integrator, and the attenuation of high frequency components is -20 dB/(10oct) (or -6 dB/oct). If we want to increase the attenuation rate, we should increase the order of the low-pass filter. That is to connect several first-order low-pass filters in series. Generally, we think that signals with frequencies lower than the cutoff frequency can pass through a low-pass filter, while other frequency components are attenuated severely.

Fig. 8.30 An active low-pass filter



The passive filter is simple, however it has several disadvantages. Firstly, as can be seen from Fig. 8.29, the passive filter does not provide gain to the input signal. Besides, the output voltage is influenced by the load resistance. To solve this problem, an active filter can be constructed with an op-amp, as shown in Fig. 8.30.

Due to virtual short, $v_- = v_+ = 0$. The current through input resistance is:

$$i = \frac{v_{in}}{R_{in}}$$

Considering virtual open and Ohm's law, the output voltage can be obtained:

$$\begin{aligned} v_{out} &= v_- - i \frac{R_f \cdot Z_C}{R_f + Z_C} \\ &= -v_{in} \frac{R_f}{R_{in}} \frac{1}{1 + j\omega R_f C_f} \end{aligned}$$

Then, the transfer function is:

$$H(j\omega) = \frac{v_{out}}{v_{in}} = -\frac{R_f}{R_{in}} \frac{1}{1 + j\omega R_f C_f} \quad (8.58)$$

The magnitude response can be derived from Eq. (8.58):

$$A(j\omega) = |H(j\omega)| = \frac{R_f}{R_{in}} \frac{1}{\sqrt{1 + (\omega R_f C_f)^2}} \quad (8.59)$$

It can be noticed from Eq. (8.59) that the gain of the active low-pass filter can be adjusted by the ratio of R_f / R_{in} . The cutoff frequency of the active low-pass filter is also the frequency corresponds to a reduction of 3 dB in magnitude. Thus the cutoff frequency is:

$$f_L = \frac{1}{2\pi R_f C_f} \quad (8.60)$$

8.4.2 High-Pass Filter

High-pass filter has the effect contrary to low-pass filter. It allows the pass of high-frequency signals and attenuates the low-frequency signals. A passive high-pass filter is shown in Fig. 8.31. It is obtained by switching the positions of capacitor and resistor of the low-pass filter.

Assuming the input and output voltages are v_{in} and v_{out} , the governing equation is:

$$v_{out} + \frac{1}{RC} \int v_{in} dt = v_{in} \quad (8.61)$$

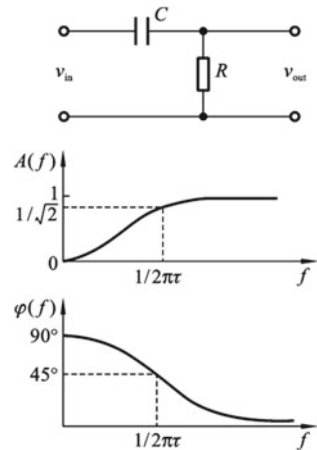
The transfer function of the passive high-pass filter can be obtained using a process similar to the one shown in Sect. 8.3.1:

$$H(j\omega) = \frac{j\omega RC}{1 + j\omega RC} \quad (8.62)$$

The magnitude and phase responses can be further obtained:

$$A(j\omega) = |H(j\omega)| = \frac{1}{\sqrt{1 + \frac{1}{(\omega RC)^2}}} \quad (8.63)$$

Fig. 8.31 A passive high-pass filter



$$\varphi(j\omega) = \arctan \frac{\text{Im}(H(j\omega))}{\text{Re}(H(j\omega))} = -\arctan\left(\frac{1}{\omega RC}\right) \quad (8.64)$$

or

$$A(f) = |H(f)| = \frac{2\pi f \tau}{\sqrt{1 + (2\pi f \tau)^2}} \quad (8.65)$$

$$\varphi(f) = -\arctan\left(\frac{1}{2\pi f \tau}\right) \quad (8.66)$$

As an example, the magnitude and phase spectra of the passive high-pass filter with $R = 100 \, \Omega$ and $C = 10 \, \mu\text{F}$ is shown in Fig. 8.32. Similarly, the cutoff frequency is given by:

$$f_H = \frac{1}{2\pi RC} \quad (8.67)$$

When $f \gg 1/(2\pi\tau)$, $A(f) = 1$ and $\varphi = 0$. Under this condition, signals can pass without attenuation, and the RC filter can be considered as an undistorted system.

An active high-pass filter can be constructed with op-amp as shown in Fig. 8.33. Due to virtual short and virtual open, the current in the feedback loop is:

$$i = \frac{v_{in}}{R_{in} - j\frac{1}{\omega C}}$$

The output voltage is:

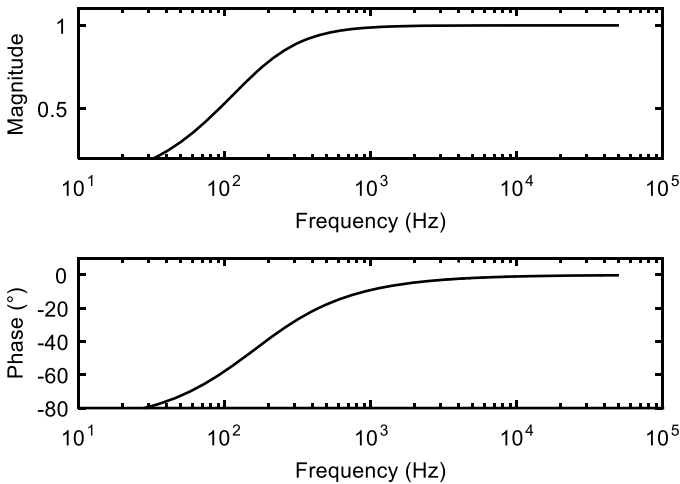
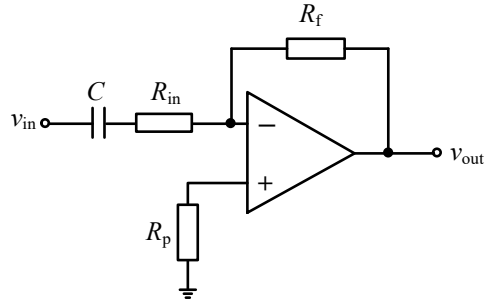


Fig. 8.32 The magnitude and phase spectra of a high-pass filter

Fig. 8.33 An active high-pass filter



$$v_{\text{out}} = -i R_f = -\frac{v_{\text{in}} R_f}{R_{\text{in}} - j\frac{1}{\omega C}}$$

Then, the transfer function is:

$$H(j\omega) = -\frac{j\omega C R_f}{1 + j\omega C R_{\text{in}}} \quad (8.68)$$

8.4.3 Band-Pass Filter

Band-pass filter is a device that allows frequencies within a certain range to pass. Band-pass filter is made by cascading a low-pass filter and a high-pass filter, as schematically shown in Fig. 8.34. The lower cutoff frequency (f_L) of the band-pass filter corresponds to the cutoff frequency of the high-pass filter while the higher cutoff frequency (f_H) of the band-pass filter corresponds to the cutoff frequency of the low-pass filter. Thus, to form a pass band, f_H must be higher than f_L . And the bandwidth is given by $f_H - f_L$.

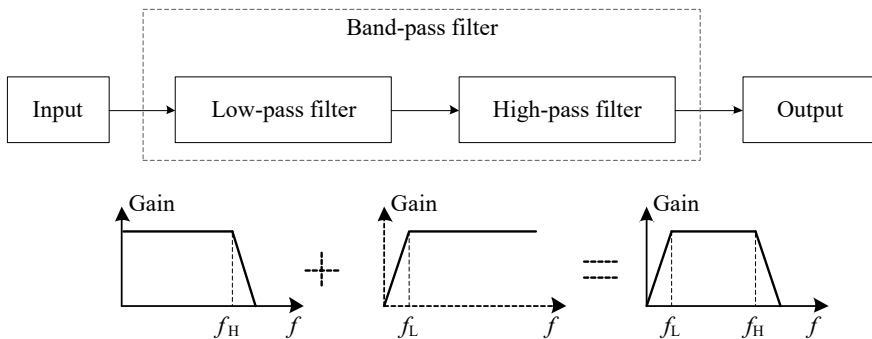
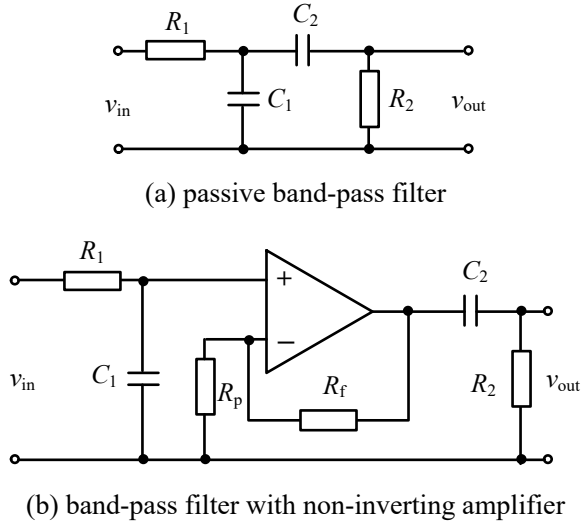


Fig. 8.34 Schematic diagrams of band-pass filter

Fig. 8.35 Band-pass filters

A basic passive band-pass filter is shown in Fig. 8.35a by cascading the low-pass filter in Fig. 8.28 and high-pass filter in Fig. 8.31. The pass band of the filter can be adjusted by the time constants τ_1 and τ_2 of the low-pass and high-pass filters. A disadvantage of this band-pass filter is that the input impedance of the high-pass filter is acting as a load influencing the characteristics of the low-pass filter. At the same time, the output impedance of the low-pass filter also influences the characteristics of the high-pass filter. To avoid these influences, a voltage follower can be inserted as a separator. More practically, as shown in Fig. 8.35b, an active non-inverting amplifier can be inserted to amplify the signal while filtering.

An active band-pass filter can be constructed by cascading the active low-pass filter in Fig. 8.30 and the high-pass filter in Fig. 8.33. More commonly, the circuit is simplified using one op-amp as shown in Fig. 8.36. Due to virtual short, $v_- = v_+ = 0$. The current through C_1 and R_1 is:

$$i = \frac{v_{in} - v_-}{R_1 - j\frac{1}{\omega C_1}} = v_{in} \frac{j\omega C_1}{j\omega C_1 R_1 + 1}$$

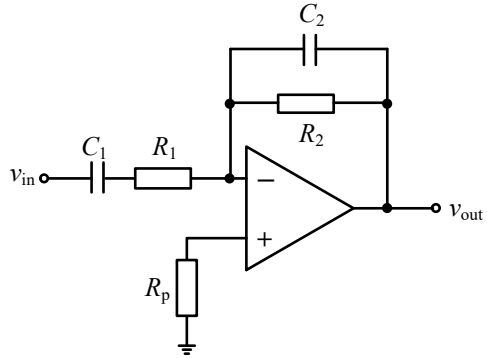
The output voltage is:

$$v_{out} = v_- - i \cdot (R_2 \parallel Z_{C2}) = -v_{in} \frac{j\omega C_1 R_2}{(1 + j\omega C_1 R_1)(1 + j\omega C_2 R_2)} \quad (8.69)$$

The transfer function can be obtained from Eq. (8.69):

$$H(j\omega) = \frac{v_{out}}{v_{in}} = -\frac{j\omega C_1 R_2}{(1 + j\omega C_1 R_1)(1 + j\omega C_2 R_2)} \quad (8.70)$$

Fig. 8.36 An active band-pass filter



The lower and higher cutoff frequencies can be obtained from Eq. (8.70):

$$f_L = \frac{1}{2\pi R_1 C_1} \quad (8.71)$$

$$f_H = \frac{1}{2\pi R_2 C_2} \quad (8.72)$$

The gain of the band-pass filter is a function of frequency, while the maximum gain at center frequency $f = (f_L + f_H)/2$ is:

$$A_v = \frac{R_2}{R_1} \quad (8.73)$$

The gain and phase shift of the band-pass filter can be plotted against frequency using Eq. (8.70). An example is shown in Fig. 8.37 with $R_1 = 1 \text{ k}\Omega$, $C_1 = 500 \text{ nF}$, $R_2 = 10 \text{ k}\Omega$, $C_2 = 1 \text{ nF}$.

8.4.4 Band-Stop Filter

Band-stop filter, which is also called notch filter, is a device that stops a range of frequency while allowing others to pass. It is commonly used to remove a special frequency component in the signal, such as the power frequency noise of 50 Hz (or 60 Hz, depending on the country). It also consists of a low-pass filter and a high-pass filter as schematically shown in Fig. 8.38. For the band-stop filter, the cutoff frequency of the low-pass filter should be lower than the cutoff frequency of the high-pass filter. A basic band-stop filter is shown in Fig. 8.39.

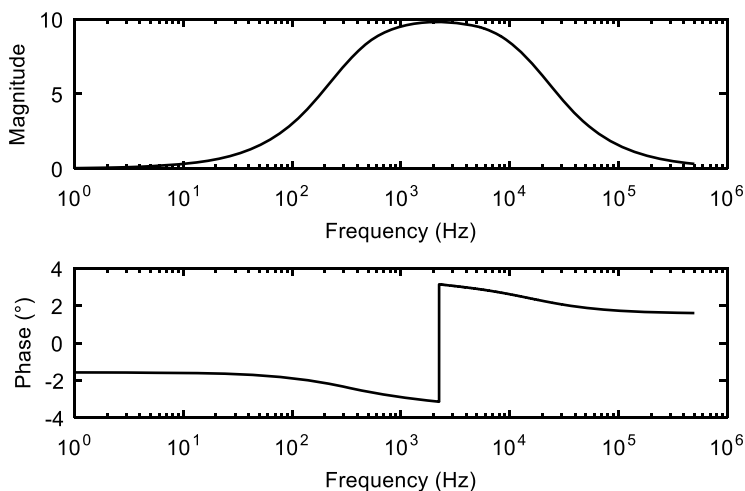


Fig. 8.37 The magnitude and phase spectra of a band-pass filter

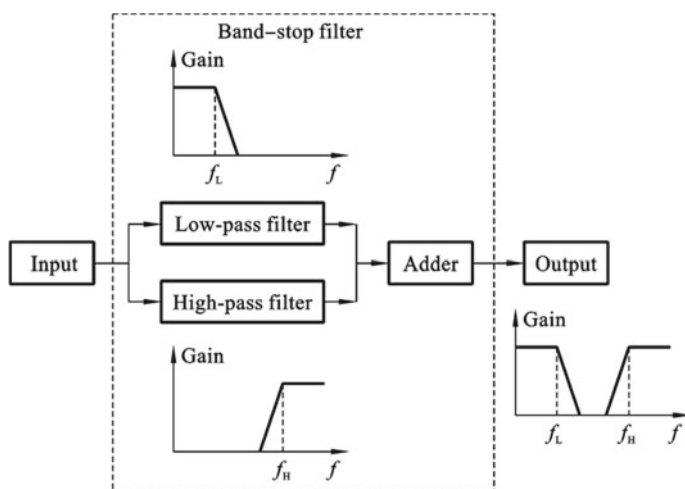


Fig. 8.38 Schematic diagram of band-stop filter

8.4.5 Comparison of Digital Filters and Analog Filters

Compared with analog filters, digital filters have the same function but different analysis methods. The mathematical model of the digital filter is a difference equation, and the operations are delay, multiplication, and addition. The components that make up the digital filter are adders, multipliers, delays, and so on. The mathematical

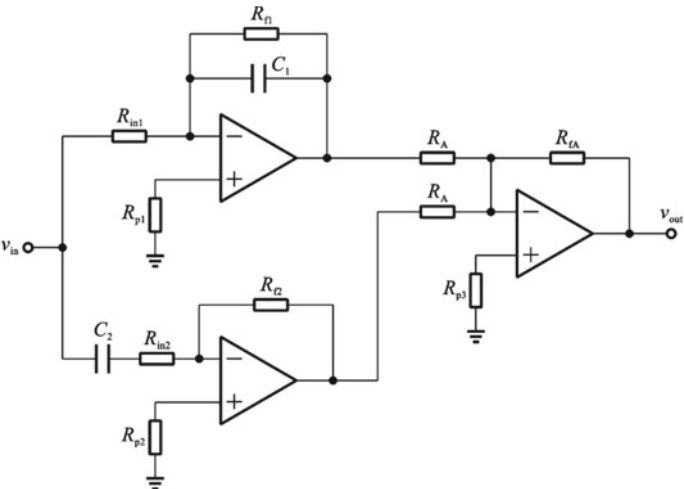


Fig. 8.39 A basic band-stop filter

model of the analog filter is a differential equation, and the operations are differentiation, integration, multiplication and addition. The components that make up analog filters are resistors, capacitors, operational amplifiers, and so on. The comparison between the two can be summarized in Table 8.1.

Digital filtering can be implemented by software or hardware. Software implementation is to compile a computer program according to the relationship between the output and the input sequence represented by the difference equation and run it on a personal computer; the hardware implementation is to connect adders, multipliers

Table 8.1 Comparison of digital filters and analog filters

Item	Analog filter	Digital filter
Input/output	Analog signal	Digital signal
System	Continuous-time	Discrete-time
Properties	Time-invariant, superposition, homogeneous	Shift-invariant, superposition, homogeneous
Mathematical model	Difference equation	Differential equation
Operations	Differentiation, integration, multiplication and addition	Delay, multiplication, and addition
Constitutions	Resistors, capacitors and op-amps	Software: program Hardware: adders, multipliers, delays
Transfer functions	$H(s) = \frac{Y(s)}{X(s)}$ (s domain) $H(\omega) = \frac{Y(\omega)}{X(\omega)}$	$H(z) = \frac{Y(z)}{X(z)}$ (z domain) $H(e^{j\omega}) = \frac{Y(e^{j\omega})}{X(e^{j\omega})}$



Fig. 8.40 Modulation and demodulation of signals

and delayers according to the diagram. The application of digital filters is similar to that of analog filters.

8.5 Modulation and Demodulation

Signal modulation is a technique in which measurement signal is used to modify certain parameters, such as amplitude, frequency or phase, of a carrier signal, so that the carrier signal carries the information of measurement signal. Signal demodulation is the inverse process of modulation, which extracts information from the carrier signal and obtains the original measurement signal. The process of signal modulation and demodulation is shown in Fig. 8.40.

8.5.1 Amplitude Modulation

1. Mathematical analysis of synchronous modulation and demodulation

In amplitude modulation, the amplitude of carrier signal changes according to the measurement signal. The process of synchronous amplitude modulation and demodulation is shown in Fig. 8.41. The modulator for amplitude modulation is a multiplier, and the demodulator consists of a multiplier and a filter.

Mathematically, amplitude modulation is equivalent to multiplying the measurement signal with carrier signal in time domain:

$$y(t) = x(t) \times c(t) \quad (8.74)$$

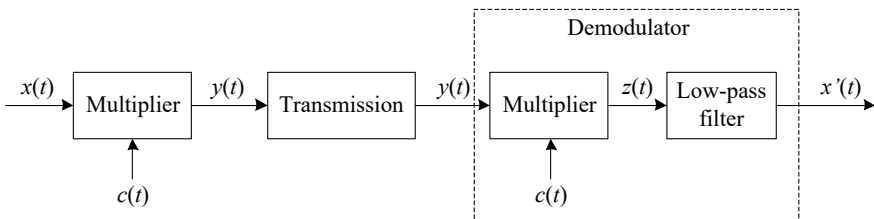


Fig. 8.41 Process of amplitude modulation and demodulation

where $x(t)$ is the measurement signal from sensor, which is also called modulating signal, $c(t)$ is the carrier signal, and $y(t)$ is the modulated signal. Usually, the carrier signal is sinusoidal, and has a frequency higher than that of the measurement signal. Substituting the carrier signal in Eq. (8.74) with a cosine signal, the following expression is obtained:

$$y(t) = x(t) \cdot A \cdot \cos(2\pi f t + \varphi) \quad (8.75)$$

Rearranging the quantities in Eq. (8.75), we get:

$$y(t) = [A \cdot x(t)] \cdot \cos(2\pi f t + \varphi) \quad (8.76)$$

From Eq. (8.76), we can notice that the amplitude of the cosine signal, $A \cdot x(t)$, changes according to the measurement signal $x(t)$. An example of the signal waveforms is shown in Fig. 8.42, in which, the measurement signal is assumed to be a sinusoidal signal

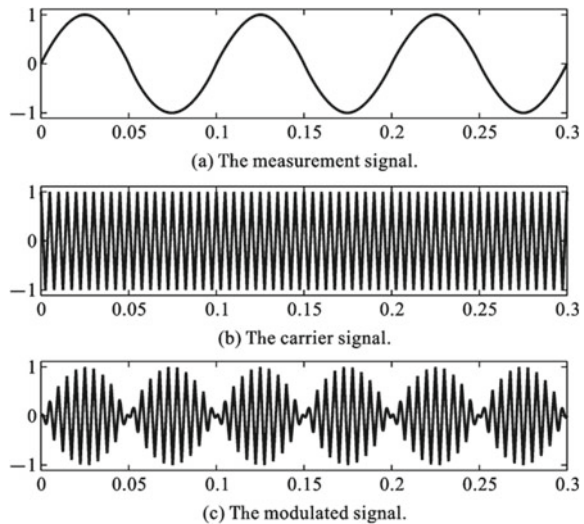
$$x(t) = \sin(2\pi \cdot 10 \cdot t) \quad (8.77)$$

and the carrier signal is assumed to be

$$c(t) = \cos(2\pi \cdot 200 \cdot t) \quad (8.78)$$

The demodulation is a two-step process. Firstly, the modulated signal in Eq. (8.75) is multiplied with the same carrier signal to get:

Fig. 8.42 Signals in the modulation process



$$\begin{aligned}
z(t) &= y(t) \cdot c(t) \\
&= x(t) \cdot A^2 \cdot \cos^2(2\pi ft + \varphi) \\
&= x(t) \cdot A^2 \cdot \frac{1 + \cos(4\pi ft + 2\varphi)}{2} \\
&= \frac{1}{2}x(t)A^2 + \frac{1}{2}x(t)A^2 \cos(4\pi ft + 2\varphi)
\end{aligned} \tag{8.79}$$

Then, with the low-pass filter, the second term in Eq. (8.79) is removed, and the output is:

$$x'(t) = \frac{1}{2}A^2x(t) \tag{8.80}$$

The output signal in Eq. (8.80) has the same waveform as the measurement signal $x(t)$, whereas it has a change in the amplitude. An amplifier with gain of $2/A^2$ can be further added to get the original signal. This demodulation method is called synchronous demodulation because the carrier signal used in the demodulation should be synchronized in phase with the one used in modulation.

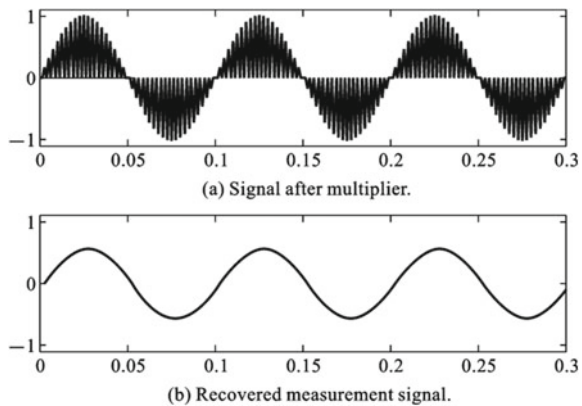
In order to demodulate the signal of the signal in Fig. 8.42c, it is multiplied with the carrier signal again, and the result is shown in Fig. 8.43a. A third order Butterworth low-pass filter with cutoff frequency of 100 Hz is designed to filter the signal in Fig. 8.43a, and the result after filtering is shown in Fig. 8.43b. The recovered measurement signal has the same waveform as the original signal shown in Fig. 8.42a, while its amplitude is only half of the original one.

2. Frequency domain analysis of synchronous modulation and demodulation

The multiplication of two signals in time domain is equivalent to the convolution of their Fourier transforms in frequency domain:

$$y(t) = x(t) \times c(t) \xleftrightarrow{F} Y(f) = X(f) * C(f) \tag{8.81}$$

Fig. 8.43 Signals in the demodulation process



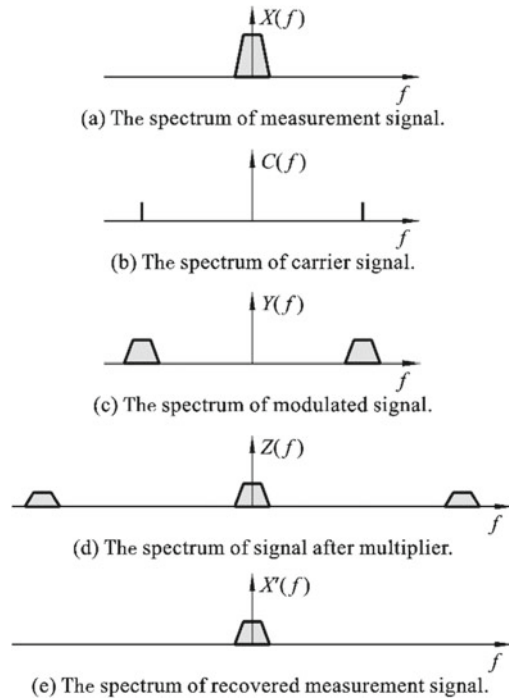
where $Y(f)$, $X(f)$ and $C(f)$ are respectively the Fourier transforms of $y(t)$, $x(t)$ and $c(t)$.

The Fourier transform of the sinusoidal carrier signal is the unit impulse function shown in Fig. 8.44b, and the Fourier transform of measurement signal is arbitrarily shown in Fig. 8.44a. The result of the convolution of a function with unit impulse function is obtained by shifting the function into the position of the unit impulse. Thus, the spectrum of the modulated signal can be obtained by the shifting and is shown in Fig. 8.44c. In the demodulation process, the modulated signal should multiply with the carrier signal again. Correspondingly, in the frequency domain, the spectrum in Fig. 8.44c should be shifted from origin to the positions of unit impulse and be added together. The result is shown in Fig. 8.44d. Finally, with the low-pass filter, the high-frequency components are removed and the resulting spectrum is shown in Fig. 8.44e. This spectrum has a change in the magnitude but it has the same waveform as the original spectrum in Fig. 8.44a.

3. Analog circuit for synchronous modulation and demodulation

Mathematically, modulation is to multiply the measurement signal with the carrier signal. Thus, the multiplier circuit in Fig. 8.13 can be used as the analog modulator. The multiplier circuit is usually simplified by the symbol shown in Fig. 8.45a. The demodulator consists of a multiplier and a low-pass filter. Thus, as shown in

Fig. 8.44 Spectra of signals in the modulation and demodulation process



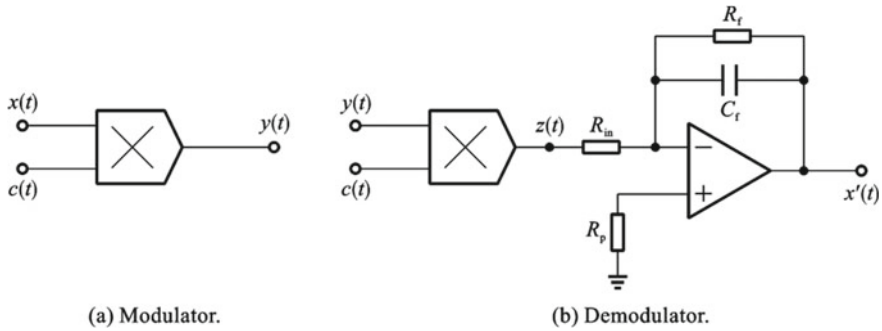


Fig. 8.45 Analog circuit for amplitude modulation and demodulation

Fig. 8.45b, it can be obtained by cascading the multiplier and the low-pass filter shown in Fig. 8.30.

4. Asynchronous amplitude modulation and demodulation

In synchronous demodulation, a carrier signal which has the same phase as the one used in modulation is required. However, the modulated signal is usually transmitted to a far-away place to be demodulated. Thus, it is difficult to get a carrier signal that is in phase with the one used in modulation, which makes the synchronous demodulation method unpractical. This problem can be solved by introducing the asynchronous amplitude modulation and demodulation, and the process is shown in Fig. 8.46.

In asynchronous amplitude modulation, a DC value is added to the measurement signal:

$$x_b(t) = x(t) + b$$

The DC value should be large enough to make all the values of the biased signal larger than zero. The modulated signal is obtained by multiplying the biased signal with the carrier signal:

$$y(t) = [x(t) + b] \cdot A \cdot \cos(2\pi ft + \varphi)$$

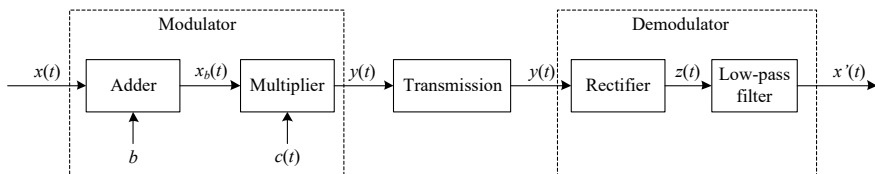


Fig. 8.46 Asynchronous amplitude modulation and demodulation

An example is shown in Fig. 8.47. The measurement signal is the same as the one in Eq. (8.77). A DC value of 1.5 is added to the measurement signal so that the biased signal in Fig. 8.47b has all values larger than zero. The carrier signal is the same as the one shown in Eq. (8.78), and the modulated signal is shown in Fig. 8.47c. The demodulator is actually an envelope detector, which consists of a rectifier and a low-pass filter. The function of the rectifier is to flip the negative part of the modulated signal into positive values, the result after rectification is shown in Fig. 8.47d. Finally, a low-pass filter is used to recover the measurement signal.

5. Distortions in amplitude modulation and demodulation

(1) Over-modulation distortion

In asynchronous modulation, a DC offset is added to the measurement signal. If the DC value is too small that the biased signal still has negative values, over-modulation distortion would occur, as shown in Fig. 8.48. In this example, the measurement signal is the same as the one in Fig. 8.47a. A bias of 0.5 is used so that there are negative values in the biased signal shown in Fig. 8.48a. In this case, the retrieved measurement signal is different from the original measurement signal.

Fig. 8.47 Signals in asynchronous amplitude modulation

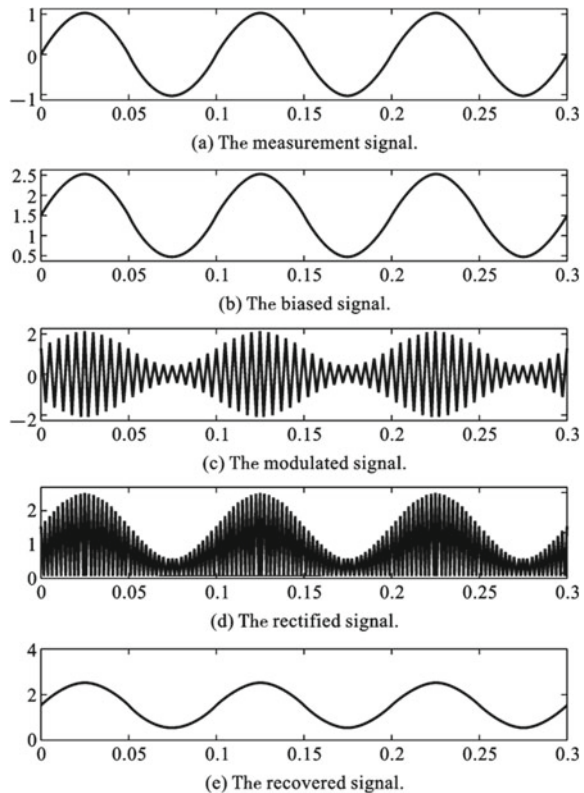
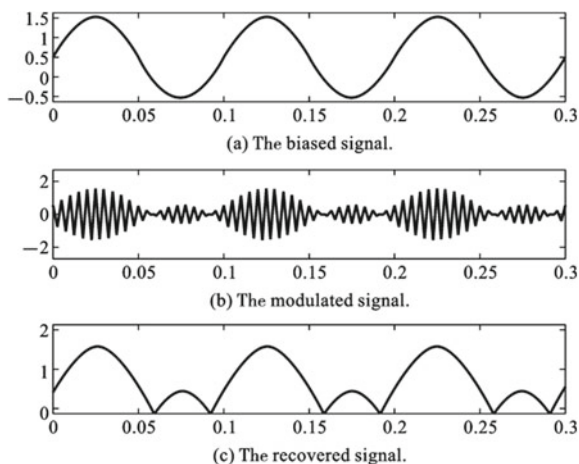


Fig. 8.48 Over-modulation

(2) Overlapping distortion

In frequency domain analysis of synchronous modulation and demodulation, we saw that the spectrum of the modulated signal has a lower sideband and an upper sideband. If the frequency of the carrier signal is small, the two sidebands would overlap as shown in Fig. 8.49. This would cause overlapping distortion to the recovered measurement signal.

As an example, the same measurement signal shown in Eq. (8.77) is used, which has a frequency of 10 Hz. The carrier signal is chosen to be a cosine signal with frequency of 20 Hz, as shown in Fig. 8.50b. The modulated signal and the recovered signal are shown in Fig. 8.50c, d respectively. It can be noticed that the recovered signal has a different waveform from the original measurement signal. While in the previous example shown in Figs. 8.42 and 8.43, when a carrier signal with 200 Hz is used, the recovered signal has the same waveform as the original one. Thus, in synchronous modulation and demodulation, the carrier signal should have a frequency that is much larger than the frequency of modulating signal.

(3) Distortion due to system characteristics

The system characteristics also influence the transmission of signal and cause distortion. During the transmission of modulated signal, no distortion would occur if the

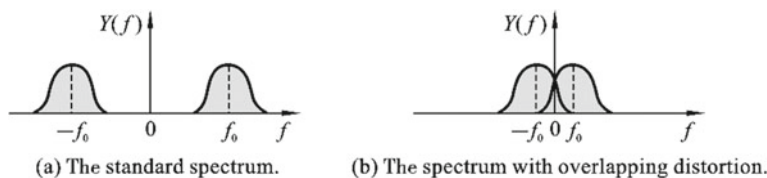
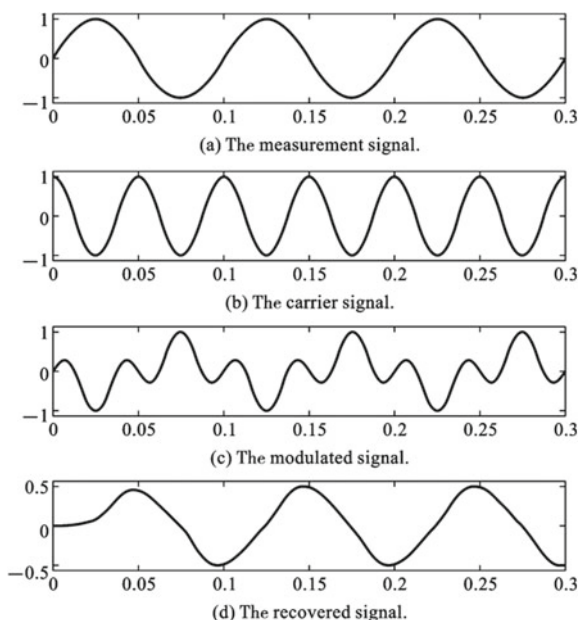
**Fig. 8.49** Spectra of modulated signal

Fig. 8.50 Overlapping distortion



signal passes a system with full bandwidth as shown in Fig. 8.51. However, if the transfer function has different magnitudes for different frequencies, then of waveform of the signal would be changed.

Example 8.3 AM Broadcast Amplitude modulation (AM) is a commonly used way to broadcast radio signals. The sound wave recorded in the radio station is multiplied with a radio-frequency carrier signal. The modulated signal is transmitted

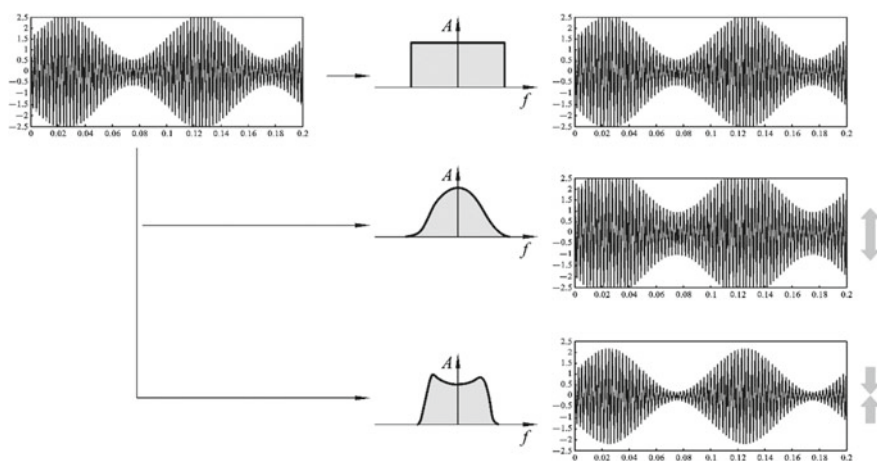


Fig. 8.51 System induced distortion

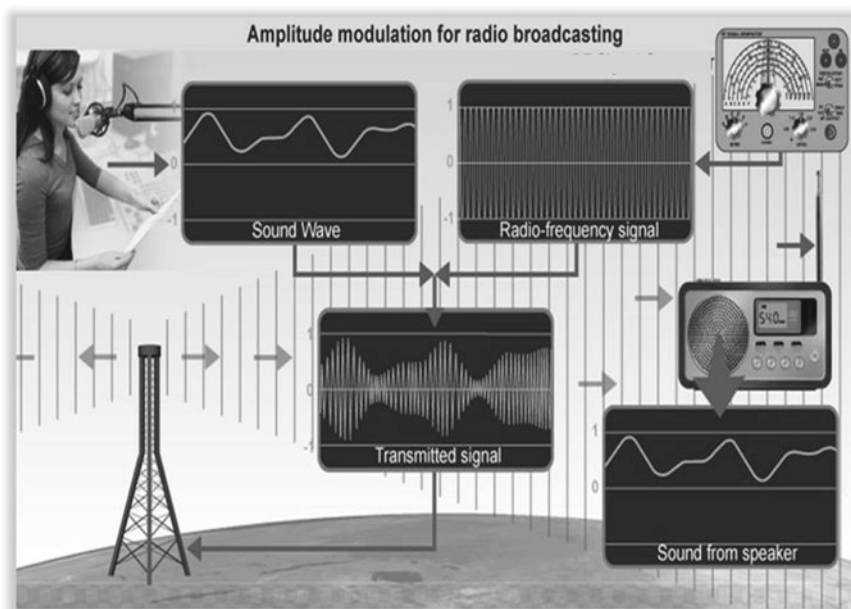


Fig. 8.52 AM broadcast

from radio station by a tower. A radio can be used to receive the broadcasting signal and demodulate it into the original sound wave, so that people at home can hear the sound recorded in the radio station (Fig. 8.52).

Example 8.4 Infrared Remote Control Infrared remote control has been widely used in both home appliances and industries, e.g. the remote control of TV set. Each function (such as power on/off, volume up/down) of the remote control is represented by a digital signal consists of series of 0 and 1. Similar to the principle of Morse code, short and long pulses are used to represent 0 and 1 in remote control. Figure 8.53 shows a signal of series “1001”, in which logical 1 is represented by a pulse of 1.2 ms, logical 0 is represented by a pulse of 0.6 ms and there are intervals of 0.6 ms to separate the pulses. The signal is multiplied with a sinusoidal carrier signal of 35 kHz to get the modulated signal. The modulated signal is transmitted to the TV set through the infrared LED in front of the remote control panel. There is an infrared receiver in the TV set to receive the infrared signal and a demodulator to recover the digital signals.

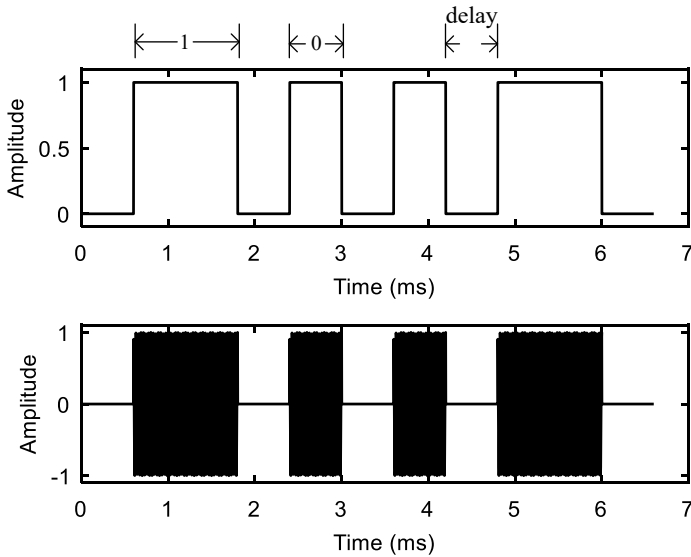


Fig. 8.53 Modulation in infrared remote control

8.5.2 Frequency Modulation and Phase Modulation

During signal transmission, the modulated signal is attenuated when passing through various media. For example, when the radio broadcasting signal passes through cloud, its amplitude changes due to attenuation. Since the signal amplitude is used for recovery in amplitude modulation, the attenuation during transmission would cause errors. Thus, the amplitude modulation has bad resistance to interferences. To avoid the interference, angle modulation can be used, which includes frequency modulation and phase modulation.

A carrier signal can be written in the following form:

$$c(t) = \cos(2\pi f_c t + \varphi_c)$$

In frequency modulation, the derivative of the carrier signal frequency is proportional to the measurement signal:

$$\frac{df_c}{dt} = k \cdot x(t)$$

Hence, the frequency modulated signal is expressed as

$$y(t) = \cos(2\pi \int df_c \cdot t + \varphi_c) = \cos\{2\pi [f_0 + k \cdot \int_{-\infty}^t x(t) dt] \cdot t + \varphi_c\} \quad (8.82)$$

In phase modulation, the phase changes according to the measurement signal while the frequency is kept constant:

$$\varphi_c = \varphi_0 + k \cdot x(t)$$

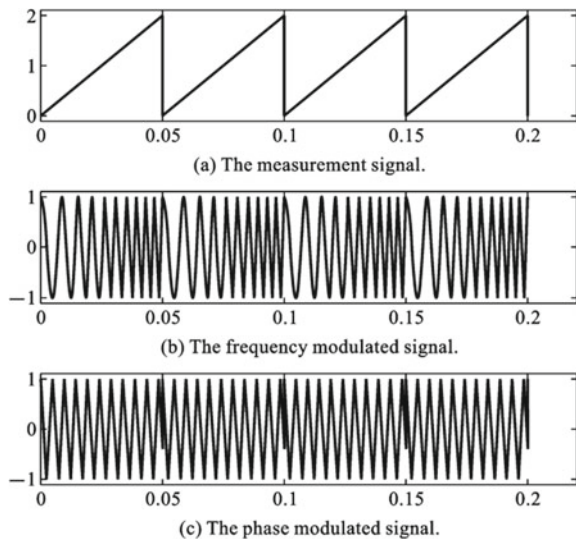
Hence, the phase modulate signal is:

$$y(t) = \cos[2\pi f_c t + \varphi_0 + k \cdot x(t)] \quad (8.83)$$

An example of frequency and phase modulation is given in Fig. 8.54. It can be seen from the figure that the frequency of the frequency modulated signal changes according to the measurement signal. The demodulation of frequency modulated signal is realized by using a zero-crossing detector, which finds the change of period. The change of frequency can be further derived and then the measurement signal can be recovered. On the other hand, a phase detector can be used to detect the phase change in the phase modulated signal. However, the change of phase in the phase modulated signal is difficult to observe except for the point where measurement signal suddenly changes. Therefore, phase modulation is commonly used for modulating digital signals only consist of zeros and ones.

The frequency and phase modulation has the advantage of better anti-interference because the frequency and phase are not easily affected by transmitting medium. Besides, both frequency and phase modulated signals have uniform amplitude, thus the modulation system can always work at peak power. However, the frequency modulation has a disadvantage of broader bandwidth. A comparison between the amplitude and frequency modulated signal and their corresponding spectra is shown in Fig. 8.55. The frequency modulated signal obviously has a broader bandwidth.

Fig. 8.54 Frequency and phase modulation



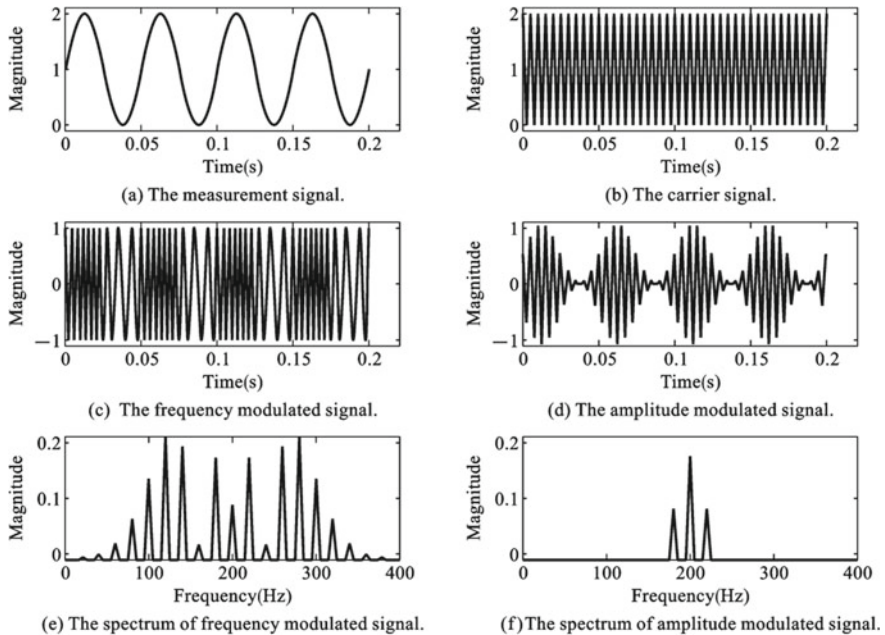


Fig. 8.55 Comparison of amplitude modulation and frequency modulation

Example 8.5 Automatic Dispatch of Train Frequency modulation can be used for the automatic dispatch of trains. When the red light is on, the station sends a signal modulated with sinusoids of 29.0 Hz, while when the green light is on, it sends a signal modulated with sinusoids of 10.3 Hz. The demodulator in the train can detect signal frequency then sends a signal to allow the train to pass or stop.

Example 8.6 Wireless Signal Transmission of a Rotating Probe In many applications, the sensor signal can be transmitted by wires directly. However, in some other measurement applications, such as magnetic flux leakage testing of axial defects in steel pipe, the probe (which contains sensors) must keep rotating around the pipe. If wires are connected to the sensors, they will intertwine with each other. Therefore, wireless transmission is required. The process of signal transmission is shown in Fig. 8.56. The measurement signal from sensor is amplified and filtered, then converted into the change of frequency by the AD654 chip. The modulated signal is used to drive an LED after power amplification. At the receiving end, a photosensitive element is used to acquire the optical signal. After amplification, the signal is demodulated by the AD650 chip.

Exercise

1. What are the types and functions of analog amplifiers?
2. How to use an operational amplifier to design an active low-pass filter?

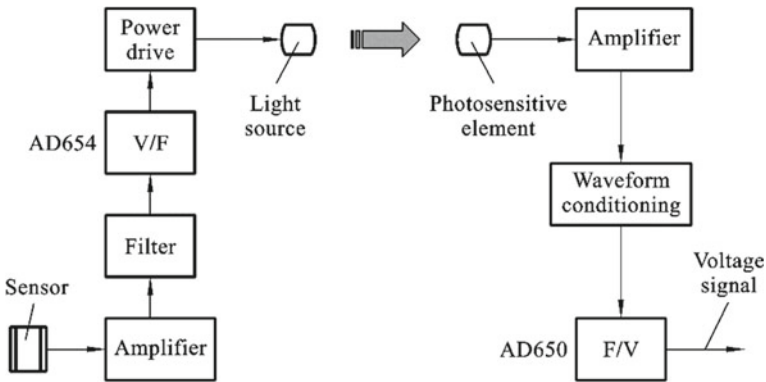
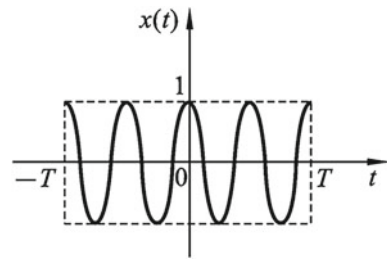


Fig. 8.56 Wireless signal transmission for rotating probe

Exercise Fig. 8.1 A cosine function modulated by a rectangular pulse



3. As shown in Exercise Fig. 8.1, the cosine function is amplitude modulated by a rectangular pulse, and its mathematical expression is

$$x(t) = \begin{cases} \cos \omega_0 t & |t| < T \\ 0 & |t| > T \end{cases}$$

Try to find its spectrum.

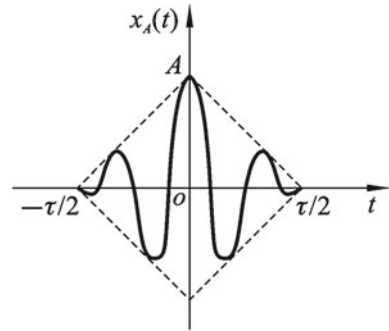
4. A cosine signal $x_2(t) = \cos \omega_0 t$ is amplitude modulated by a triangular pulse $x_1(t)$ as shown in Exercise Fig. 8.2. The Fourier transform of the triangular pulse is assumed to be

$$X_1(\omega) = \frac{A\tau}{2} \sin^2(\omega\tau/4)$$

Try to find the spectrum of the modulated signal.

5. Describe the working principle of the modulated DC amplifier.

Exercise Fig. 8.2 A cosine signal modulated by a triangular pulse



6. Analyze the spectral characteristics of the AM signal, and explain why the excitation frequency of the bridge of strain gauge is much higher than the working frequency of the strain gauge.
7. For a modulated signal with a polarity change, how should the signal be demodulated to reflect the polarity change of the original signal.

Chapter 9

Characteristics of Measurement System



9.1 Overview of Measurement System

Measurement system is a unit that fulfills the task of measurement, it usually includes: sensors, signal conditioning module and signal analysis/display module as shown in Fig. 9.1. In the measurement system, sensor is used to convert physical quantity into electric quantity; signal conditioning module is used to amplify and filter the electric signal, which facilitates the subsequent signal transmission and processing; the analysis and display module is to extract information from the signal and display it.

The characteristics of measurement system can be divided into two categories: static characteristic and dynamic characteristic. The static characteristic is to describe the system performance when the measured physical quantity remains unchanged or changes very slowly. Whereas, the dynamic characteristic is used to describe the relationship between the input and output of a system when the measured physical quantity changes rapidly.

9.2 Static Characteristics of Measurement Systems

The static characteristics are used to describe the performance of a measurement system when the input is static or quasi-static. The characteristics are obtained by calibration, in which a standard value is used as input and the corresponding output is extracted. The static characteristics mainly include the following parameters.

1. Sensitivity

The output of a measurement system changes by Δy in response to the change of input Δx . Their ratio is defined as the sensitivity:

$$S = \frac{\Delta y}{\Delta x} \quad (9.1)$$

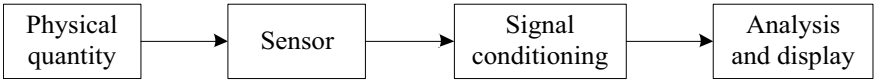


Fig. 9.1 Schematic diagram of a measurement system

For example, if the output voltage of a digital thermometer is 1 V at 10 °C, and it is 1.2 V at 11 °C, then the sensitivity of the thermometer is 0.2 V/°C. Sensitivity is a very important parameter in measurement system, it determines if small change in the input can be measured or not. Usually, the larger the sensitivity the better. However, a measurement system with large sensitivity is more likely to pick up the ambient noise. Thus, the signal to noise ratio should also be considered.

2. Accuracy

Accuracy is used to measure the ability of a measurement system to output a value that matches the true value. The difference between the measured result and the true value is called systematic error. For example, if two electronic scales are used to measure a standard weight of 2 kg. One outputs a value of 2.01 kg and the other one outputs 2.02 kg, then the first one is said to have better accuracy. If the systematic error is a constant, it can be compensated.

3. Resolution

Resolution of a measurement system refers to the smallest amount of change that can be detected. For example, if a distance measurement system has a resolution of 1 mm, a change of 0.5 mm will not cause any change in the output.

4. Repeatability

Repeatability refers to the ability of a measurement system to provide the same reading under repeated tests of same input and same measuring conditions. Repeatability is also called precision. To understand the difference between accuracy and precision, an example is given in Table 9.1 with two electronic scales measuring a standard weight of 2 kg. As we can see, the measurement results of Scale 1 fluctuate more than that of Scale 2. The fluctuation can be mathematically described by the standard deviation σ . Since Scale 2 has smaller standard deviation, it has better precision. Although Scale 1 has worse precision, the mean of measurement results is more close to the true value. Thus, Scale 1 has better accuracy.

5. Nonlinearity

Table 9.1 Measurement results of two electronic scales (units are in kg)

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	μ	σ
Scale 1	2.06	2.08	1.93	1.93	2.05	1.97	1.98	2.00	0.06
Scale 2	2.09	2.08	2.09	2.08	2.07	2.07	2.08	2.08	0.01

If the sensitivity of the measurement system is not a constant within the measurement range, then the relationship between input and output is no longer linear. The solid line shows the input–output relationship of a measurement system, and the dashed line is the result of linear fitting. Nonlinearity is defined as:

$$e_{NL} = \frac{\Delta y_L}{y_R} \quad (9.2)$$

where Δy_L is the largest deviation from the input–output curve to the fitted line, y_R is the output range. For example, a measurement system with variable-distance capacitive sensor is nonlinear. To reduce the nonlinearity, we can reduce the measurement range and use differential sensors (Fig. 9.2).

6. Hysteresis error

For some measurement systems, the output and input do not have one to one correspondence, i.e. the output value depends not only on the input but also on the history of input change. Hysteresis error is defined as:

$$e_H = \frac{\Delta y_H}{y_R} \quad (9.3)$$

where Δy_L is largest difference between two output values of a same input during the increasing and decreasing of input, y_R is the output range (Fig. 9.3).

7. Threshold

Fig. 9.2 Nonlinearity

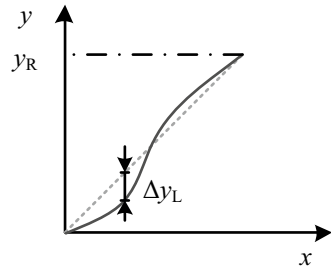
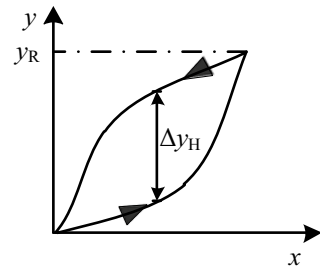


Fig. 9.3 Hysteresis



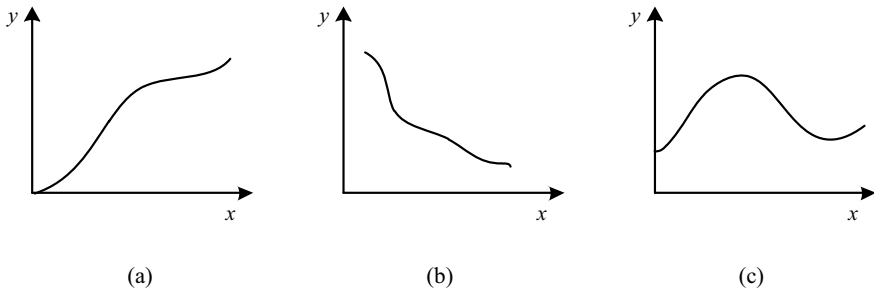


Fig. 9.4 Monotonicity. **a** Monotonous increasing, **b** monotonous decreasing and **c** non-monotonic

Threshold refers to the input range where the output keeps constant while input is increasing or decreasing from zero, which is the resolution near the input value of zero.

8. Input/output range

Input range is the range between the minimum and maximum measurable input values. And the range between corresponding outputs is called output range.

9. Monotonicity

If a measurement system is monotonous, its output always increases with the increase of input as shown in Fig. 9.4a or it always decrease as shown in Fig. 9.4b. If the relationship between input and output is as shown in Fig. 9.4c, then the measurement system is non-monotonic.

9.3 Dynamic Characteristics of Measurement Systems

When the physical quantity to be measured changes rapidly, the output will also be a rapidly changing quantity. Then, the measurement system is called a dynamic system, and the output is related to the input by the dynamic characteristics: impulse response function and transfer function.

9.3.1 Transfer Function and Frequency Response Function

In time domain, the output is the convolution of input and impulse response function:

$$y(t) = h(t) * x(t) \quad (9.4)$$

The impulse response function can be obtained by inputting a unit impulse to the system and measuring the output signal. In frequency domain, the spectrum of output is:

$$Y(f) = H(f) \cdot X(f) \quad (9.5)$$

where $H(f)$ is the system transfer function and $X(f)$ is the spectrum of input. The transfer function of a system can be obtained by calculating the spectra of input and output and taking their ratio. The transfer function is usually an array of complex numbers, which can be written as:

$$H(j\omega) = H_R(\omega) + j \cdot H_I(\omega) \quad (9.6)$$

The amplitude spectrum and phase spectrum can be further derived from the transfer function:

$$A(\omega) = \sqrt{H_R^2(\omega) + H_I^2(\omega)} \quad (9.7)$$

$$\varphi(\omega) = \arctan\left(\frac{H_I(\omega)}{H_R(\omega)}\right) \quad (9.8)$$

The amplitude spectrum in Eq. (9.7) describes the ratio of output amplitude and input amplitude, while the phase spectrum in Eq. (9.8) describes the phase shift between input and output signals. An example is given in Fig. 9.5.

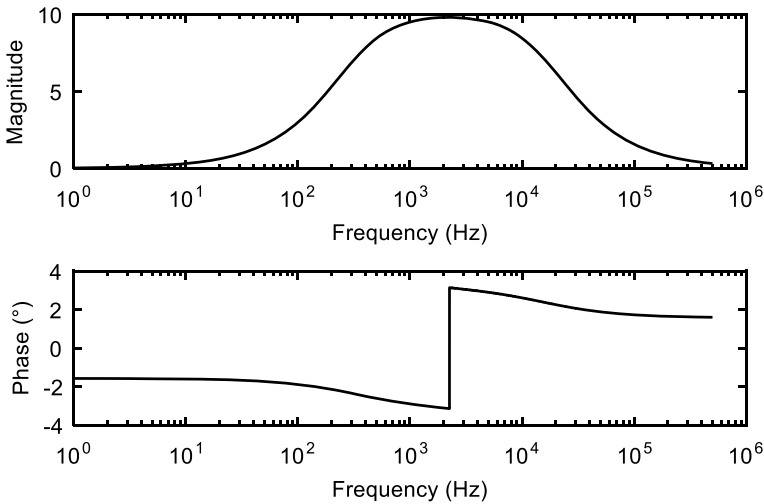


Fig. 9.5 The magnitude and phase spectra of a system with band-pass filter

There are three types of problems in system analysis. The first one is system identification, where the input $x(t)$ and output $y(t)$ are known (measurable), and the system characteristics $h(t)$ or $H(f)$ need to be determined. The second one is inverse solution, where the system characteristics $h(t)/H(f)$ are known, and the input $x(t)$ needs to be determined given a measured output $y(t)$. The third one is prediction, where input $x(t)$ and system characteristics $h(t)/H(f)$ are known, and the output $y(t)$ needs to be estimated.

9.3.2 Linear Measurement System

The ideal measurement system should have a single-valued, determinant input–output relationship. Among them, the linear relationship between output and input is the best. In static measurement, this linear relationship of the measurement system is always desirable, but it is not necessary, because curve correction or output compensation technology can be used for non-linear correction in static measurement. In dynamic measurement, the measurement system should try to be a linear one, not only because at present only linear systems can be mathematically analysis comprehensively, but also because it is quite difficult to make nonlinear corrections in dynamic measurement. It is impossible for some actual measurement systems to maintain complete linearity within a larger working range, so linear processing can only be done within a certain working range and within a certain allowable range of error. Strictly speaking, the actual measurement system always has non-linear factors. For example, many electronic devices are non-linear, but the measurement system is often treated as a linear system in engineering, which can simplify the problem while maintaining sufficient accuracy.

1. Definition of linear system

If the relationship between system input and output can be described by a constant coefficient linear differential equation

$$\begin{aligned} & a_n y^n(t) + a_{n-1} y^{n-1}(t) + \cdots + a_1 y(t) + a_0 \\ & = b_m x^m(t) + b_{m-1} x^{m-1}(t) + \cdots + b_1 x(t) + b_0 \end{aligned} \quad (9.9)$$

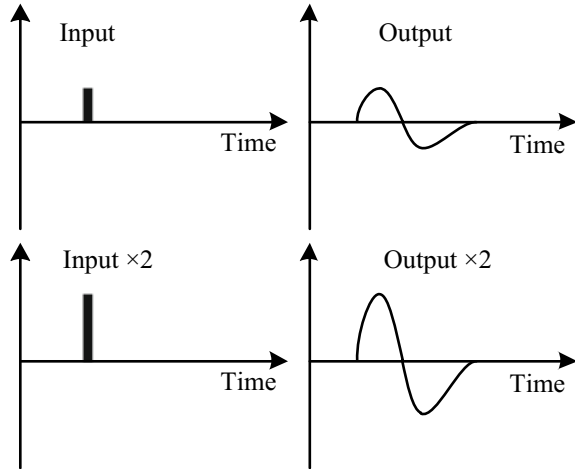
then the system is called a linear system. In frequency domain, the transfer function satisfies:

$$H(f) = \frac{Y(f)}{X(f)} = \frac{b_m f^m + b_{m-1} f^{m-1} + \cdots + b_1 f + b_0}{a_n f^n + a_{n-1} f^{n-1} + \cdots + a_1 f + a_0}$$

Usually, the measuring instruments used in engineering can be approximated as a linear system within the measuring range.

2. Properties of linear system

Fig. 9.6 Property of proportionality



(1) Proportionality

As illustrated in Fig. 9.6, the property of proportionality indicates that if the output for an input $x(t)$ is $y(t)$, then the output of the input $kx(t)$ should be $ky(t)$.

(2) Superposition

The property of superposition indicates that, if the output for the inputs $x_1(t)$ and $x_2(t)$ are respectively $y_1(t)$ and $y_2(t)$, then the output of the input $x_1(t) + x_2(t)$ is $y_1(t) + y_2(t)$, as shown in Fig. 9.7.

(3) Differentiation

The property of differentiation indicates that if the output for an input $x(t)$ is $y(t)$, then the output of the input $x'(t)$ should be $y'(t)$.

(4) Integration

The property of integration indicates that if the output for an input $x(t)$ is $y(t)$, and the initial condition is zero, then the output of the input $\int x(t)$ should be $\int y(t)$.

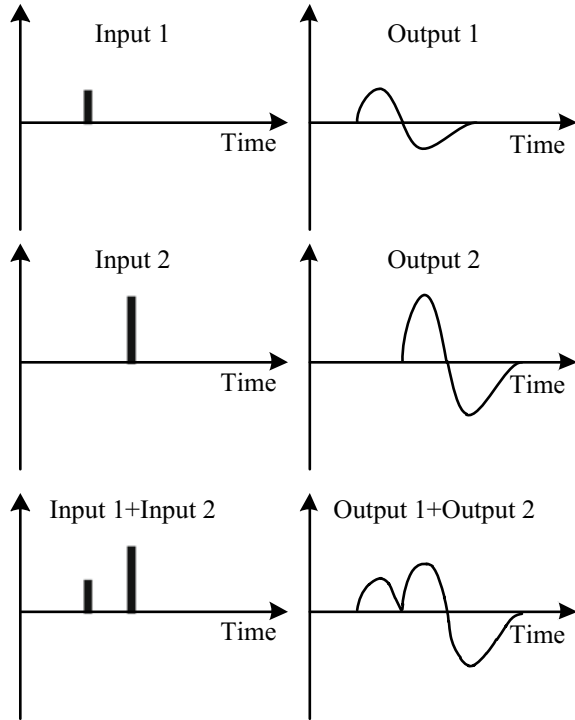
(5) Frequency retentivity

The property of frequency retentivity indicates that the output of a linear measurement system always has the same frequency as the input.

3. Methods of measuring dynamic characteristics

The dynamic characteristics of a measurement system can be determined by inputting a standard signal and measuring the corresponding output. The commonly used input signals are impulse function, step function, ramp function, sinusoidal function and white noise.

Fig. 9.7 Property of superposition



(1) Impulse function

The most intuitive way to determine dynamic characteristics is to input an impulse function to the measurement system. In time domain, if the impulse function is used as input, the output is directly the impulse response function that we want to determine:

$$y(t) = h(t) * \delta(t) = h(t)$$

The transfer function can be obtained by taking the Fourier transform of the impulse response. An example of determining dynamic characteristics with impulse function is shown in Fig. 9.8. The advantage of using impulse function as input is that the testing is simple and intuitive, while the disadvantage is that the input energy of the impulse is relatively low, thus, sometimes, we cannot get a measurable output signal.

(2) Swept frequency sinusoids

If the input is a single frequency (f_0) sinusoid, the output will also be a sinusoid with the same frequency f_0 while having changes in amplitude and phase. By measuring the amplitude and phase changes, the value of transfer function at this particular frequency $H(f_0)$ can be obtained according to Eq. (9.6) to Eq. (9.8). Repeating the

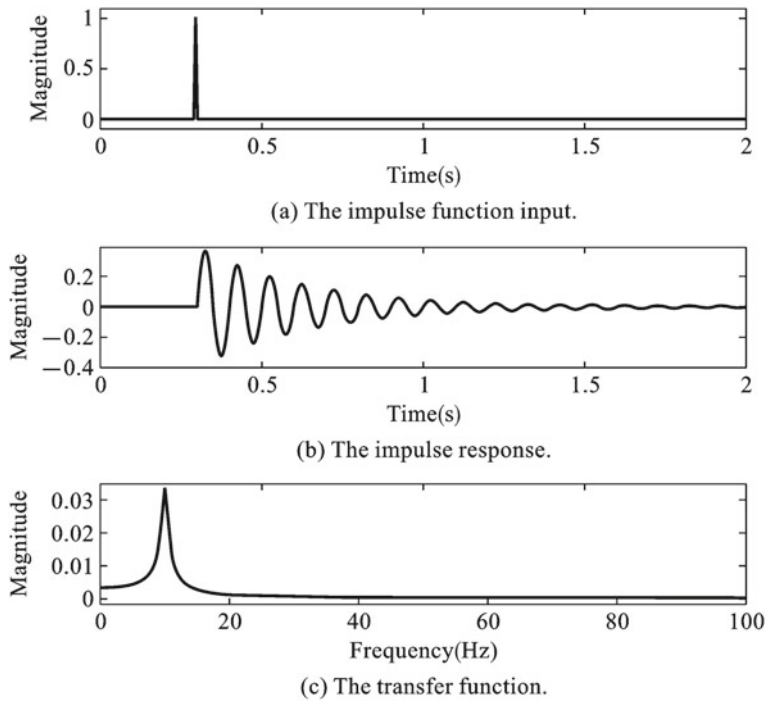


Fig. 9.8 Determining dynamic characteristics with impulse function

process by sweeping the frequency of input sinusoid, the whole transfer function and the corresponding spectra can be obtained. The advantage of using swept frequency sinusoid is that the input signal has higher energy. At the same time, it has the disadvantage of low efficiency because the test should be repeated several times with different frequencies.

(3) White noise

To avoid repeating the tests with various frequencies, we can use a signal containing all the desired frequencies as the input signal. White noise is a signal with random values in time domain and flat spectral density in frequency domain.

9.4 Characteristics of Typical Linear Measurement Systems

The relationship between input and output of a linear system can be expressed by the general equation shown in Eq. (9.9). Among the linear systems, the most common ones are zero order system:

$$a_1 y(t) + a_0 = x(t) \quad (9.10)$$

first order system:

$$a_2 \frac{dy(t)}{dt} + a_1 y(t) + a_0 = x(t) \quad (9.11)$$

and second order system:

$$a_3 \frac{d^2 y(t)}{dt^2} + a_2 \frac{dy(t)}{dt} + a_1 y(t) + a_0 = x(t) \quad (9.12)$$

Different systems will cause different distortions to the signal. For the task of measurement, it is desired to have a system without distortion, i.e. the output waveform is the same as the input waveform (Fig. 9.9). Mathematically, the output and input should satisfy:

$$y(t) = A \cdot x(t - t_0) \quad (9.13)$$

By applying Fourier transform to Eq. (9.13), the undistorted condition can be examined in frequency domain:

$$Y(f) = A \cdot e^{-j2\pi f t_0} \cdot X(f)$$

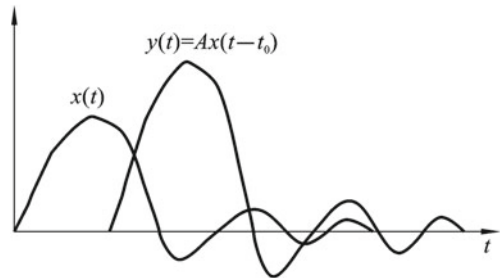
The amplitude spectrum and phase spectrum of the system can be further derived:

$$A(f) = A = \text{constant}$$

$$\varphi(f) = 2\pi t_0 f$$

If a measurement system satisfies the above-mentioned time-domain or frequency-domain characteristics, i.e. its amplitude-frequency characteristic is a constant, and the phase-frequency characteristic has a linear relationship with frequency, then

Fig. 9.9 Undistorted measurement



the system is said to be an accurate or undistorted measurement system. The measurement implemented by the system will be accurate and distortion-free.

9.4.1 Zero Order System

For a zero order system, the relationship between input and output satisfies the zero-order differential equation. The following equation represents a zero order system:

$$y(t) = k \cdot x(t) \quad (9.14)$$

Applying Fourier transform to Eq. (9.14), the transfer function can be determined:

$$H(f) = \frac{Y(f)}{X(f)} = k \quad (9.15)$$

As we can see from Eq. (9.14), the output changes with the input instantaneously, hence the zero order system has no delay. Besides, Eq. (9.14) obviously satisfy the condition shown in Eq. (9.13), thus the zero order system is an undistorted measurement system. The zero order system also has infinite frequency bandwidth due to the constant transfer function shown in Eq. (9.15). For example, a displacement measurement system with potentiometer (Example 7.3) is a zero order system.

9.4.2 First Order System

For a first order system, the relationship between input and output is a first order differential equation shown in Eq. (9.11). Applying Laplace transform to Eq. (9.11), the transfer function can be obtained:

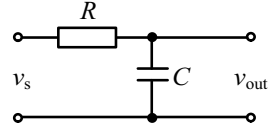
$$H(s) = \frac{Y(s)}{X(s)} = \frac{k}{\tau s + 1}$$

Example 9.1 A Measurement System Based on RC Circuit Figure 9.10 shows a system which connects a RC low-pass filter to the sensor. According to Ohm's law, the following equation can be obtained:

$$iR + v_{\text{out}} = v_s$$

where i is the current, v_s is sensor voltage, v_{out} is the measured voltage. The current is related to charge and voltage across the capacitor as:

Fig. 9.10 A measurement system with RC circuit



$$i = \frac{dq}{dt}$$

$$q = C v_{out}$$

Thus, the time domain equation for the measurement system is:

$$RC \frac{dv_{out}}{dt} + v_{out} = v_s \quad (9.16)$$

which satisfies the equation of first order system in Eq. (9.11). The analysis in frequency domain has already been done in Sect. 8.3.1 and the transfer function is given as:

$$H(j\omega) = \frac{1}{j\omega RC + 1}$$

Example 9.2 Measure temperature with mercury thermometer When a mercury thermometer is inserted into a liquid with constant temperature of T_L , the measured temperature T_M changes with time by absorbing heat from (or releasing heat to) the liquid. The heat stored in the mercury is given by:

$$Q = mCT_M$$

where m is the mass of mercury, C is the heat capacity, thus

$$kS(T_L - T_M) = \frac{dQ}{dt} = mC \frac{dT_M}{dt}$$

where k is the equivalent conduction coefficient, S is the surface area of mercury thermometer. Re-organizing the equation, we get:

$$\frac{mC}{kS} \frac{dT_M}{dt} + T_M = T_L$$

Therefore, the mercury thermometer is also a first order measurement system.

1. Characteristics of first order system in time domain

The output of a first order system always has a delay when compared with input. For the RC circuit system described by Eq. (9.16), suppose the input sensor voltage is a

step function that increase from zero to E_s at $t = 0$, then the output voltage can be obtained by solving Eq. (9.16), and it is given as:

$$v_{\text{out}} = E_s \left(1 - e^{-\frac{t}{RC}} \right)$$

The changes of input and output are shown in Fig. 9.11. The measured voltage increases gradually and reaches the equilibrium value of E_s in the end. The delay is measured by the time constant τ , which is the time when the output reaches $(1 - 1/e)E_s$. For the RC circuit system, the time constant is related to resistance and the capacitance: $\tau = RC$.

A ramp function can also be used as the input to measure the time constant of a first order system. Under ramp input, the output will finally become a line parallel to the ramp input as given by Fig. 9.12. Extending the parallel line to the horizontal axis, then the intersection point is the time constant.

2. Characteristics of first order system in frequency domain

The amplitude and phase of the output of a first order system depends on the input frequency. Taking the R–C circuit in Fig. 9.10 as an example, it works as a low-pass filter which stops the high frequency signal. In frequency domain, the most important parameter is the cutoff frequency, $f_c = 1/2\pi\tau$, which is the frequency when the gain is -3 dB. A more comprehensive analysis has already been done in Sect. 8.3.1.

Fig. 9.11 Step response of RC circuit

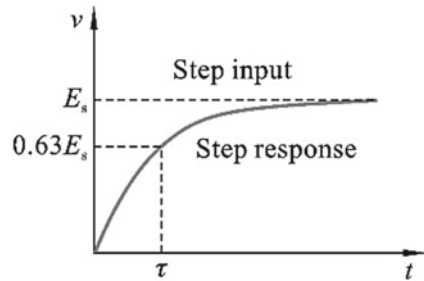
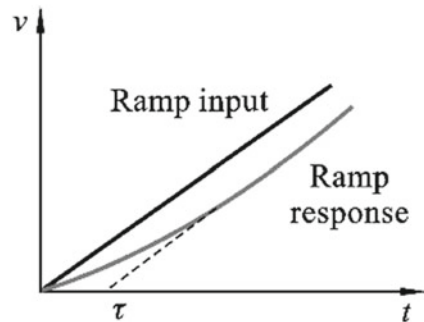


Fig. 9.12 Ramp response of RC circuit



9.4.3 Second Order System

For a second order system, the relationship between input and output is a second order differential equation shown in Eq. (9.12). Applying Laplace transform to Eq. (9.12), the transfer function of a second order system can be obtained:

$$H(s) = \frac{Y(s)}{X(s)} = \frac{1}{a_2 s^2 + a_1 s + a_0} \quad (9.17)$$

Defining the following parameters:

$$k = \frac{1}{a_0}$$

$$\xi = \frac{1}{2} \sqrt{\frac{a_1}{a_0}}$$

$$\omega_n = \sqrt{\frac{a_0}{a_2}}$$

where k is called static stiffness, ξ the damping coefficient and ω_n the natural frequency, then Eq. (9.17) can be re-written as:

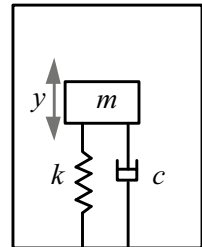
$$H(s) = \frac{k \omega_n^2}{s^2 + 2\xi \omega_n s + \omega_n^2} \quad (9.18)$$

Example 9.3 A Spring-mass System with Damping A spring-mass system with damping is schematically shown in Fig. 9.13. The mass moves under external force. According to Newton's second law

$$ma = F - c \frac{dy}{dt} - ky$$

Substituting the relationship between acceleration and displacement, we get:

Fig. 9.13 A spring-mass system with damping



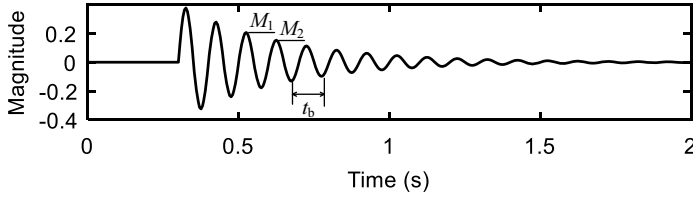


Fig. 9.14 Unit impulse response

$$m \frac{d^2 y}{dt^2} + c \frac{dy}{dt} + ky = F$$

Thus the spring-mass system with damping is a second order system.

1. Characteristics of second order system in time domain

The simplest way to measure the characteristics of a second order system is to input a unit impulse function and measure the corresponding output. A unit impulse response is shown in Fig. 9.14. The natural frequency can be calculated by:

$$\omega_n = \frac{2\pi}{t_b}$$

where t_b is the time difference between two neighboring peaks. The damping coefficient can be determined by:

$$\xi = \frac{\omega_n}{2\pi} \ln\left(\frac{M_1}{M_2}\right)$$

where M_1 and M_2 are the amplitudes of two neighboring peaks.

2. Characteristics of second order system in frequency domain

The relationship between output and input in frequency domain can be obtained by substituting $j\omega$ into Eq. (9.18):

$$H(j\omega) = \frac{k\omega_n^2}{-\omega^2 + 2j\xi\omega_n\omega + \omega_n^2} = \frac{k}{1 - \left(\frac{\omega}{\omega_n}\right)^2 + 2j\xi\left(\frac{\omega}{\omega_n}\right)} \quad (9.19)$$

The magnitude response can be plotted according to Eq. (9.19). An example is shown in Fig. 9.15 with different damping coefficient. When the damping coefficient is less than 0.7, resonance occur as indicated by the peaks in the spectrum. Conversely, if the spectrum is measured during experiment, the system characteristics can be determined from the spectrum. The natural frequency can be found by the frequency corresponding to the peak in the spectrum, and the damping coefficient is determined by:

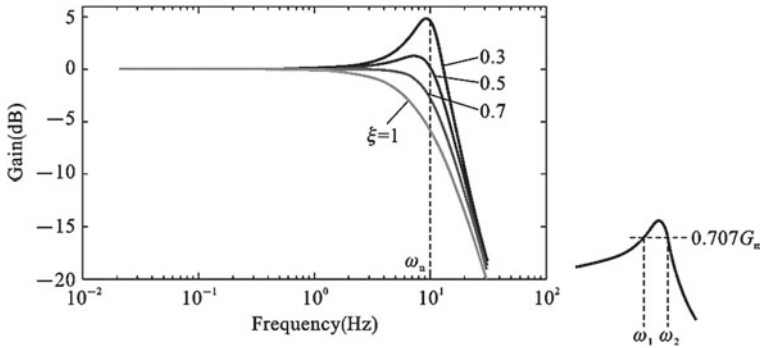


Fig. 9.15 Magnitude response of second order system

$$\xi = \frac{\omega_2 - \omega_1}{\omega_n}$$

where ω_1 and ω_2 are the frequency correspond to 0.707 times the maximum gain.

Exercise

1. What is the static characteristic of the measurement system? What are the main indicators of static characteristics?
2. What is the dynamic characteristic of the measurement system? What are the main indicators of dynamic characteristics?
3. What are the conditions for the measurement system to achieve undistorted measurement?
4. What is dynamic error? In order to reduce dynamic errors, what measures can be taken in the first and second-order measurement systems?
5. What is the physical meaning of frequency response? How is it obtained? Why does it reflect the ability of the system to measure arbitrary signals?
6. Try to explain the reason why the damping ratio of the second-order measurement system should be 0.6 ~ 0.7.
7. It is known that the natural frequency of a second-order system sensor is $f_0 = 20$ kHz, and the damping ratio $\xi = 0.1$. If the output amplitude error of the system is required to be less than 3%, try to determine the operating frequency range of the sensor.
8. Frequency response of a measurement system $H(j\omega) = \frac{1}{1+0.05j\omega}$, if a periodic signal $x(t) = 2 \cos 10t + 0.8 \cos(100t - 30)$ is input to the system, try to find the response $y(t)$.
9. A sensor has differential equation of $30 \frac{dy}{dt} + 3y = 0.15x$, where y is output voltage in mV, x is input temperature in $^{\circ}\text{C}$, try to find the time constant and the sensitivity of the sensor.
10. A force sensor is a typical second-order system with natural frequency $f_0 = 1000$ Hz of and damping coefficient of $\xi = 0.7$. If it is used to measure a

Exercise Table 9.1 The calibration data of a pressure sensor

Pressure/MPa	System output/mV					
	First test		Second test		Third test	
	Increasing input	Decreasing input	Increasing input	Decreasing input	Increasing input	Decreasing input
0.00	−2.74	−2.72	−2.71	−2.68	−2.68	−2.67
0.02	0.56	0.66	0.61	0.68	0.64	0.69
0.04	3.95	4.05	3.99	4.09	4.02	4.11
0.06	7.39	7.49	7.42	7.52	7.45	7.52
0.08	10.88	10.94	10.92	10.88	10.94	10.99
0.10	14.42	14.42	14.47	14.47	14.46	14.46

sinusoidal force of 600 Hz, what is the ratio between output and input amplitudes and what is the phase shift between them.

11. The frequency response function of a second-order measurement system is

$$H(\omega) = \frac{1}{1 - \left(\frac{\omega}{\omega_n}\right)^2 + 0.5j\left(\frac{\omega}{\omega_n}\right)}$$

Input a signal $x(t) = \cos(\omega_0 t + \frac{\pi}{2}) + 0.5 \cos(2\omega_0 t + \pi) + 0.2 \cos(4\omega_0 t + \frac{\pi}{2})$ to this system. Assume $\omega_0 = 0.5\omega_n$, try to find the time-domain response of the input.

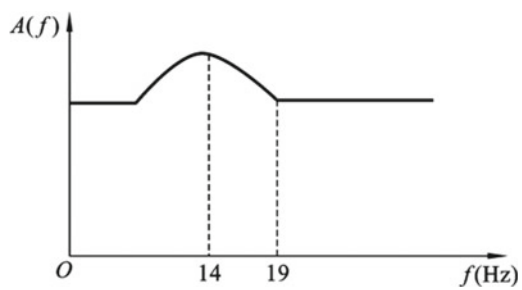
12. The calibration data of a pressure sensor is listed in Exercise Table 9.2. Find the linearity, hysteresis error and repeatability of the sensor.
13. A measurement device has the amplitude-frequency characteristic shown in Exercise Fig. 9.16. The phase shift is 75° when $\omega = 125.5$ rad/s and 180° when $\omega = 150.6$ rad/s, if the system is used to measure the following signals

$$x_1(t) = A_1 \sin 125.5t + A_2 \sin 150.6t$$

$$x_2(t) = A_3 \sin 626t + A_4 \sin 700t$$

Can the system measure x_1 and x_2 without distortion?

Exercise Fig. 9.1 The amplitude-frequency characteristic of a measurement device



Chapter 10

Computerized Measurement System



10.1 Overview of Computerized Measurement System

Since the 1970s, due to the development of large-scale integrated circuit technology, the development of computers has entered the era of microcomputers. Microcomputer has the characteristics of powerful function, small size, low power consumption, high cost performance, etc., which makes it more and more closely integrated with measurement technology. At the same time, the development of communication, network, micro-electromechanical technology and new sensor technology has continuously given new content to computerized testing technology and promoted the continuous development of measurement technology.

In a computerized measurement system, computer is used as the main part of the instrument to process and display the measurement signal. A computerized measurement system is schematically shown in Fig. 10.1. Sensor is used to convert the measured physical quantity into electric signal. Signal conditioning module is used to amplify and filter the acquired analog signal. In the signal acquisition, A/D converter is used to convert the analog signal into digital signal. Finally, a computer is used to analyze and display the digital signal. The word “computer” is a general term to describe the electronic devices such as desktop, laptop, pad and mobile phone.

Compared with traditional analog or digital instruments, the main advantages of computerized test systems are:

1. It is able to perform complex analysis and processing on the signal, such as FFT-based time domain and frequency domain analysis and vibration mode analysis.
2. It can realize high-precision, high-resolution and high-speed real-time analysis and processing. The software is used to correct the non-linear error caused by the sensor and the measurement environment. High-bits A/D converter and high-precision clock control can improve the precision and resolution of the analysis results.

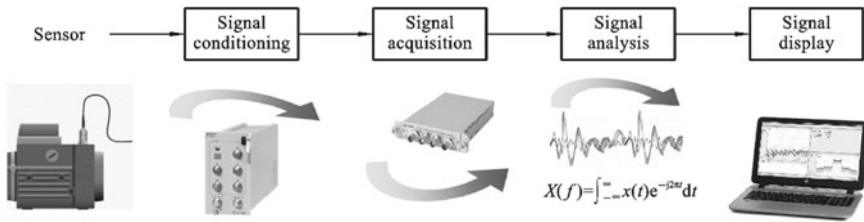


Fig. 10.1 A computerized measurement system

3. It has reliable and stable performance and it is convenient for maintenance. The computerized measurement instrument is composed of hardware and software. Mass-produced hardware guarantees high reliability and stability, easy maintenance, and good reproducibility of software operation.
4. It has the ability to output information in various forms. Various graphs and charts can visually display the analysis results. The storage of information facilitates review of analysis results. The digital communication can realize remote monitoring and remote measurement.
5. It is multifunctional. Users can expand processing functions to meet various requirements.
6. It is able to fulfill the task of automatic measurement and self-monitoring. The automatic measurement program can self-check the instrument and repair some faults.

Nowadays, computerized measurement and analysis instruments are dominant in various fields of measurement. Computerized measurement equipment is also called intelligent equipment, but it can only be regarded as primary intelligent equipment at present. The computerized measurement instrument is composed of a microcomputer, plug-in hardware and acquisition and analysis software. Insert the ADC card into the expansion slot of the microcomputer, or use a general single-chip integrated ADC to call the acquisition, analysis and processing software, then the measurement and analysis can be performed. This measurement method is usually called Computer Aided Testing (CAT), which was proposed in the 1960s. With the emergence of some high-performance ADC and DAC pre-processing modules and special measurement and analysis software, various data acquisition instruments and analyzers based on personal computers have been produced. In the late 1980s, the performance of microcomputers was greatly improved, and the successful application of general-purpose software development platforms for testing and analysis gave rise to the emergence of virtual instruments.

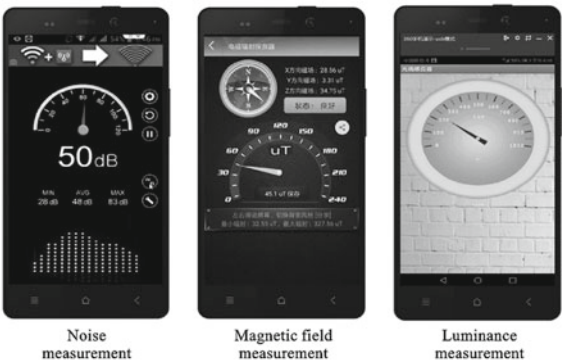
Example 10.1 A Computerized Monitoring System for Milling Spindle Vibration Monitoring To monitor the vibration of milling spindle, accelerometer, microphone and camera are used as sensor to acquire the vibrational information. The acquired analog signals are transmitted to computers after A/D conversion. The digital signals are analyzed in the computer and finally the results are displayed on the monitor (Fig. 10.2).



Fig. 10.2 Milling spindle vibration monitoring

Example 10.2 Virtual Instrument APP on Mobile Phone The simplest way for us to experience computerized measurement system is to download some specific APPs on our mobile phone. There are many sensors integrated in mobile phones, such as the magnetic field sensor, brightness sensor and microphone etc.. These sensor are designed for other purposes, e.g. the magnetic field sensor is used for navigation, the brightness sensor is used to measure the environmental luminance and automatically control the brightness of screen, the microphone is used to capture our sound during phone calls. Usually, the measured data is directly used for subsequent control or signal transmission, and we cannot see the measured data. However, if we go to the APP store to download some specific APPs, then we can see the measured data, such as the measured magnetic field in x , y and z axis. Then, we can use our mobile phone as a virtual instrument to measure some physical quantities (Fig. 10.3).

Fig. 10.3 Virtual instrument APP on mobile phone



10.2 Development of Measurement Instrument

Usually, the measurement instrument can be categorized into four generations. The first generation is based on the vacuum tube technology; the second generation is based on the transistor (integrated circuit) technology, the third generation is based on the digital technology and the fourth generation is based on the virtual instrument technology. The first and second generations are based on analog circuit and the instrument only processes analog signals. In the third and fourth generation, digital signal processing techniques are widely used to process the acquired measurement signal.

1. Vacuum tube instrument

The vacuum tube instrument is made of vacuum tube. A vacuum tube has three basic poles. One of the poles is called cathode, which is where the electrons are emitted, another one is called screen, which is the outermost metal plate of the vacuum tube. The screen is connected to a positive voltage, attracting the electrons emitted from the cathode. The other pole is gate, which is fixed between the cathode and the screen, controlling the flow of electrons. The vacuum tube is very large, thus the instrument made of vacuum tube has very large size.

2. Transistor instrument

Transistor refers to electronic components made of semiconductor materials, including diodes, transistors, field effect transistors, thyristors, etc.. With the integrated circuit technology, many transistors and wires can be integrated into one chip, thus the size of instrument is reduced. In the transistor instrument, analog circuits are widely used to amplify and filter the measured signals.

3. Digital instrument

The main difference between the digital instrument and previous generations is that the digital instrument has an A/D converter. After the A/D conversion, a specialized computer can be used to process the digital signals. Because a specialized computer is used, more complicated signal processing algorithms can be incorporated to the measurement system. Besides, with a computer, the signal can be directly displayed on the monitor.

4. Virtual instrument

Both virtual instrument and digital instrument use computers to process the digital signal. The difference is that the computer used in digital instrument is a specialized computer, whose only function is to process the measured data. While in virtual instrument, a general computer, i.e. our personal computer or laptop, is used. The virtual instrument usually has a lower price because the general computer has much larger number of sales, hence the price is lower. However, the general computer does not include A/D converters, therefore, an external A/D converter is required for the A/D conversion.

10.3 Computerized Measurement Instrument

10.3.1 Virtual Instrument

The concept of virtual instrument was proposed by National Instrument in 1986. A virtual instrument mainly includes computer, virtual instrument software, external measurement hardware and data bus as shown in Fig. 10.4. In a virtual instrument, the software is the key of the entire system to process signal, while the hardware is only used for signal acquiring and transmission.

A comparison of conventional instrument and virtual instrument is made in Fig. 10.5. For the conventional instruments, there is an independent identity that we can see. However, the virtual instrument is consisted of several parts and there is no independent instrument. The term “virtual” in virtual instrument means that there is no independent identity for the instrument.

1. Characteristics of virtual instrument

Virtual instrument is a measurement and control system composed of computer hardware and software for data analysis, process communication and graphical user interface display. It is a instrument system operated by a computer. Compared with traditional instruments, virtual instruments have the following characteristics:

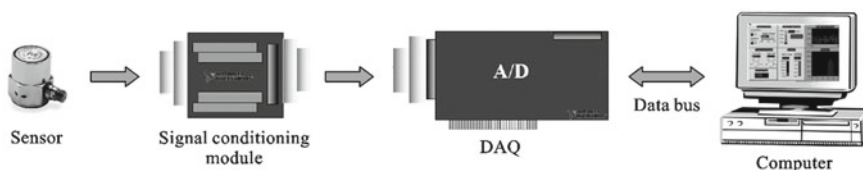


Fig. 10.4 Schematics of a virtual instrument

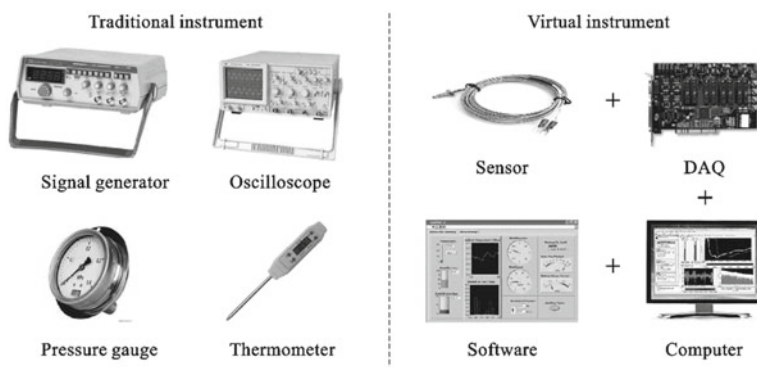


Fig. 10.5 Comparison of conventional instrument and virtual instrument

- (1) Virtual instrument users can flexibly define the functions of the instrument according to their own needs, and can form a variety of instruments through the combination of different functional modules, without being limited to the specific functions provided by the instrument manufacturer.
- (2) The virtual instrument concentrates all instrument control information in the software module, and can display the collected data, analysis results and control process in a variety of ways. This transfer of key parts further increases the flexibility of virtual instruments.
- (3) The key of virtual instrument lies in software, and the limitation of hardware is small, so it is easier to connect with other instruments. Moreover, virtual instruments can be easily connected to the network, peripherals and other applications, and can also use the network for multi-user data sharing.
- (4) The virtual instrument can edit the data directly in real time, or transmit the data to the memory or printer through the computer bus. This solves the problem of data transmission on the one hand, and makes full use of the storage capacity of the computer on the other hand, so that the virtual instrument has almost unlimited data capacity.
- (5) The virtual instrument utilizes the powerful graphical user interface (GUI) of the computer. Users can directly perform various analyses and processing on measurement data in real time through software programming or using existing analysis software.
- (6) The price of virtual instruments is low, and its software-based architecture greatly saves development and maintenance costs.

A comparison is further made in Table 10.1.

Table 10.1 Comparison of virtual instruments and traditional instruments

Traditional instrument	Virtual instrument
Functions are defined by the instrument manufacturer	Functions are defined by the users
Very limited connection with other equipment	Application-oriented system structure, which can be easily connected to the network, peripherals and other applications
Small graphical interface, manual reading, small amount of information	Able to display comprehensive information of measured signal
Data cannot be edited	Data can be edited, stored and printed
Hardware is the key component	Software is the key component
Expensive	Cheap (1/10–1/5 of traditional instrument)
Closed system, fixed function, low expansibility	Open functional blocks based on computer technology that can form a variety of instruments
Slow technology update (cycle 5–10 years)	Fast technology update (cycle 1–2 years)
High development and maintenance costs	Based on the software system, the development and maintenance costs are greatly reduced

The advantages of virtual instruments in terms of performance are as follows:

- (1) High measurement accuracy and good repeatability. The advent of embedded data processors allows the establishment of mathematical models of some functions, such as FFT and digital filters, so separate analog hardware that may drift over time and requires periodical calibration is no longer needed.
- (2) The measurement speed is high. To measure several characteristics of the input signal (such as signal frequency and rise time), only a quantized data block is needed, and the signal characteristics to be measured can be calculated by the data processor. This method combines multiple tests to shorten the measurement time. In the traditional instrument system, the signal must be connected to a certain instrument to measure various parameters, which is greatly affected by the difference in cable impedance, instrument calibration and correction factor.
- (3) Switches and cables are reduced. Since all signals have a common quantization channel, it allows various measurements to use the same calibration and correction factor. In this way, the complex switch matrix and signal cables can be reduced, and the signal does not have to be switched to multiple instruments.
- (4) The system takes a short time to set up. All common modules support the same common hardware platform. When a new function is to be added to the measurement system, it is only necessary to add software to perform the new function or add a common module to expand the measurement range of the system.
- (5) The measurement function is easy to expand. Because the instrument function can be generated by the user, it is no longer hidden in the hardware and cannot be changed. In order to improve the performance of the measurement system, it is convenient to add a common module or replace a module without purchasing a completely new system.

2. Hardware system of virtual instrument

The system composition of virtual instrument includes computer, virtual instrument software, hardware interface. Hardware interfaces include data acquisition cards, IEEE488, GPIB interface cards, serial/parallel ports, plug-in instruments, VXI controllers, and other interface cards. Data acquisition card is the most commonly used form of virtual instrument, which has the characteristics of flexibility and low cost, and is used for A/D conversion and signal transmission. The commonly used data acquisition card is the PCI data acquisition card, which is inserted into the PCI slot of the computer. The advantage is that the data transmission rate is high, but the disadvantage is that the case needs to be opened for installation. PXI (PCI eXtensions for Instrumentation) data acquisition card is specially designed and optimized for the instrument. Other data acquisition card includes USB data acquisition card, RJ45 data acquisition card, and WiFi data acquisition card. These data acquisition cards have the advantage of convenience; however, their transmission rate is relatively low.

3. Software system of virtual instrument

The software structure of a complete virtual instrument system is generally divided into 4 layers.

(1) Measurement management

The user uses the program developed by the virtual instrument manufacturer to build his own set of test instruments. This is one of the advantages of the virtual instrument. Users can easily build their own measurement instruments according to their needs.

(2) Application development layer

Users can use the software development tools provided by the manufacturer (such as NI's LabVIEW software LabWindows/CVI software, etc.) for development to extend the original functions of the instrument.

(3) Instrument driver layer

An instrument driver is a collection of software programs that complete the control and communication. It is responsible for handling the communication and control for a specific instrument. It conceals the underlying complex hardware operations, and encapsulates the details of complex instrument programming, provides a simple function interface for users to use the instrument. The user calls the instrument driver in the application program to perform the operation and design of the instrument system, which simplifies the user's development work. The instrument driver is developed by the manufacturer, and there are different driver interfaces for different types of instruments. In order to provide users with convenient and easy-to-use instrument drivers, well-known instrument companies such as Tektronix and Hewlett-Packard have established the VXIPlug&Play system alliance and introduced the VISA (Virtual Instrument Software Architecture) standard.

(4) I/O bus driver layer

The I/O interface software is located between the instrument device (ie I/O interface device) and the instrument driver. It is a low-level software that completes direct access to the instrument register and provides information for the instrument driver. It is the basis for the realization of the virtual instrument system.

The most commonly used application development software is LabVIEW, which is a graphical programming platform developed by National Instrument. As shown in Fig. 10.6, the software uses a flowchart programming method instead of coding, which greatly facilitates and simplifies programming. In the front panel, the collected signal can be displayed, and many graphic controls can be inserted to control the display.

DRVI is another virtual instrument platform mainly used for teaching and experiments. The platform was developed by our team. It adopts the idea of software bread-board, in which various components can be inserted to realize rapid programming (Fig. 10.7).

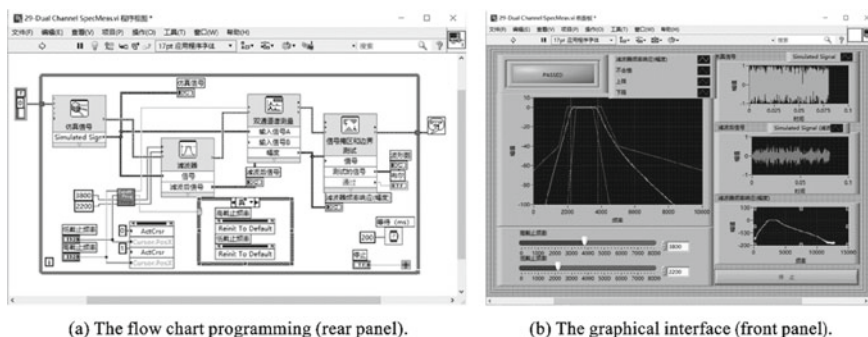


Fig. 10.6 Comparison of data acquisition cards

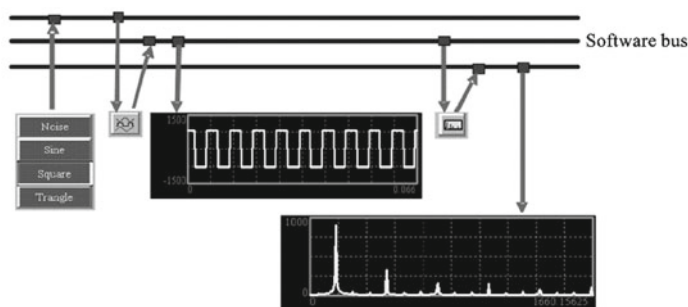


Fig. 10.7 DRVI platform

10.3.2 Network Based Measurement Instrument

In today's information society, the network has penetrated into various fields of industry, scientific research and daily life as a tool for information exchange. The development of network technology provides an opportunity for the development of network based instruments. A network based instrument is an collection of all hardware and software elements that can be remotely operated at any time and at any place to obtain measurement information. It consists of basic network hardware, application software and multiple communication protocols. Network based instruments are connected together through a network, and data can be exchanged between each other to realize data sharing. The measurement data can be transmitted to a remote place or the cloud through the network, and analyzed using remote instruments. The network based instrument can be used in the distributed control system of the production enterprise. It is distributed in different positions of the system for distributed measurement, and then the data is transmitted to the control center through the network. The control center can control the measurement process in a

remote place, which greatly improves productivity. Nowadays, network based instruments are developing rapidly, and Agilent Inc. has successfully launched network based oscilloscopes and network based logic analyzers. In addition, networked flow meters and networked sensors have also been invented. In the field of electric energy measurement, the application of remote meter reading systems is becoming more and more wide used. The electric power department can complete the acquisition and monitoring of remote meter readings through telephone lines or power lines. The electric energy meter with remote communication function used in this system is a kind of network based instrumentation.

Networked instruments have gone beyond the scope of traditional independent instruments, and are no longer a simple combination of traditional independent instrument. Based on PCs and workstations, it forms a practical measurement and control system by building a network to improve production efficiency and share information resources. In a sense, computers and modern instruments have been mutually inclusive, and computer networks are universal instruments.

At present, computer networks (represented by the Internet) have developed rapidly. With the expansion of network channel capacity, network speed no longer becomes an obstacle to network applications. Using the existing Internet facilities, networked sensors have been applied to the distributed measurement and control system, which simplifies system construction and equipment maintenance, reduces costs and improves system functions. With the development of the measurement and control network, the interconnection technology of the measurement and control network and the information network will also be improved.

1. Features of networked instruments

(1) Instrument information sharing

Under the environment of networked instrument, the measured object can transmit the measured data (information) to the remote measurement equipment or high-end computerized instrument for analysis and processing through the network to realize sharing of measurement information. In addition, data can also be transmitted to the original site through an instrument with network transmission function.

(2) Low cost and high efficiency

The measurement and control system built by the Internet and networked instruments can better integrate resources and reduce the cost of the system. In addition, the use of networked instruments can undoubtedly significantly increase the utilization rate of various complex equipment, effectively reduce the manpower and financial investment in monitoring, measurement and control work. It can also shorten the cycle of measurement, and increase the satisfaction of customers with measurement needs.

Once the traditional measurement instrument or system is combined with the network, it forms a networked instrument. Just like the remote measurement carried out by telecom service operators today, it can be obtained from any place on the earth and at any time. Measurement information needed anywhere. The development

of instrumentation and modern measurement technology are prerequisite for the emergence of the concept of networked instruments. The establishment of the concept of networked instrumentation will help people clarify the research and development strategy of instrumentation as soon as possible.

2. The working mode of networked instruments

According to different measurement requirements, there are mainly three modes of networked instruments that are common in practical applications: networked instruments based on the Client/Server model, networked instruments based on the Browser/Server model, and networked instruments based on mixed Client/Server and Browser/Server.

Client/Server (abbreviated as C/S) model was proposed and implemented by several computer companies led by Sybase in the United States in the 1980s. Later, it developed rapidly and gradually penetrated in various fields of computer applications, this architecture and network mode was once considered an ideal mode in the design of equipment remote condition monitoring and fault diagnosis instruments (Fig. 10.8).

The Browser/Server (abbreviated as B/S) structure is a multi-layer C/S structure. This mode does not require the installation of client software. Only standard browsers such as Internet Explorer, Firefox, etc. are required on the client. The use of Web technology only needs to develop and maintain server-side applications, which greatly reduces the management and maintenance of the system. All applications on the server can be executed on the client through a Web browser, which unifies the user interface (Fig. 10.9).

C/S mode and B/S mode have their own advantages and disadvantages. In actual development, B/S and C/S coexist and cooperate with each other. They are often used

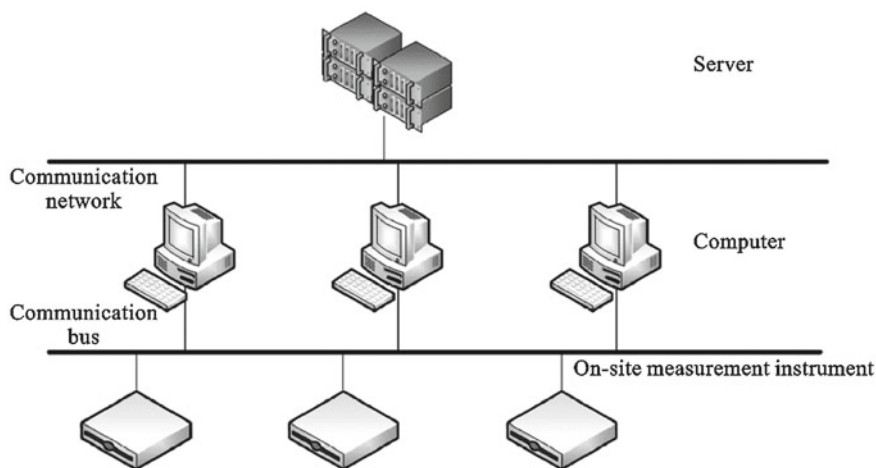


Fig. 10.8 Schematic diagram of typical networked instrument based on C/S mode

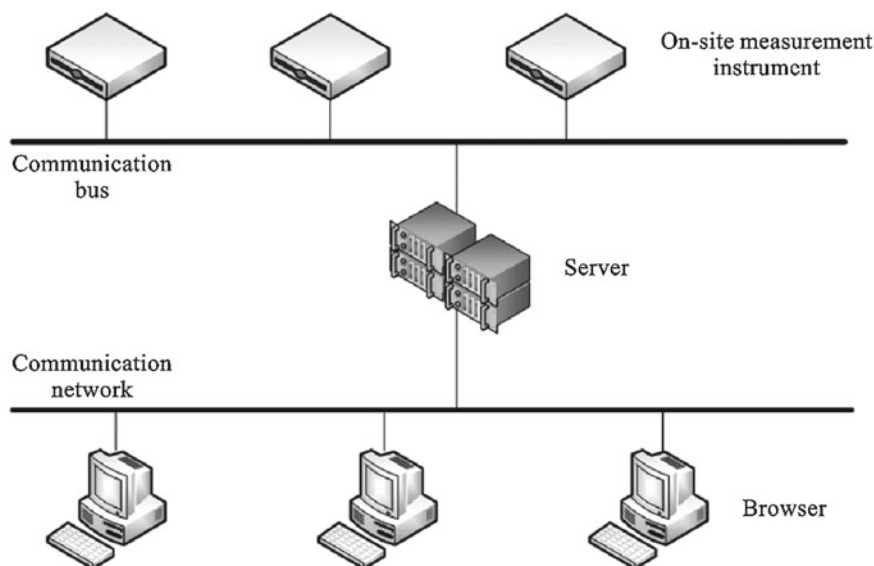


Fig. 10.9 Schematic diagram of typical networked instrument structure based on B/S mode

together in the system structure, namely C/S and B/S hybrid mode. The instrument use the C/S and B/S modes to realize the sub-functions.

3. Typical applications of networked instruments

(1) Networked traffic monitoring system

Oil, natural gas, and water are all important resources of the country. China has realized projects such as the South-to-North Water Diversion and West-East Gas Transmission through a large number of buried pipelines. The networked flow monitoring system is to set a monitoring point every tens or hundreds of kilometers in the pipeline, measure the flow of each monitoring point in the pipeline through ultrasonic sensors, and send the monitoring data to the monitoring sites for comprehensive analysis. The flow rate at the monitoring point determines whether there is a pipeline leak.

(2) Networked electric energy meter

The electric energy meter is an indispensable instrument for monitoring the electricity consumption of each household. The traditional electric energy meter is installed in a residential building and the meter is read from house to house by the staff to obtain the user's monthly electricity consumption. The networked electric energy meter can upload the electric energy meter data to the power supply management department via cable or radio. The staff can monitor the power consumption in the area in real time, and can also feedback the data to the power generation department to adjust the power generation in time.

(3) Networked environmental monitoring instrument

Changes in the climate and environment are closely related to people's daily lives. How to effectively monitor them is the focus of environmental monitoring research. The emergence of networked environmental monitoring equipment has made up for the shortcomings of low efficiency and waste of manpower in previous environmental monitoring. Networked environmental monitoring instruments can dynamically monitor environmental changes and provide reliable information and data for atmospheric monitoring, weather forecasting, natural disaster prevention, aerospace, etc. In addition, the environmental monitoring of small spaces has always been the focus, such as the real-time monitoring of the environment of waiting rooms, theaters, conference rooms, offices, wards, etc.

For example, as an important water resource, surface water has a great impact on people's livelihood. Effective detection of surface water environment is of great significance to the protection of water resources. Figure 10.10 shows a networked water quality automatic monitoring system, which mainly includes water collection unit, water distribution and pretreatment unit, control unit, analysis unit, sample retention unit, auxiliary unit, etc. It is an intelligent, standardized, process-oriented, and traceable quality control system. It has automatic processes such as water collection, pretreatment, analysis, quality control, cleaning, data collection, transmission, etc., which has improved the shortcomings of low monitoring efficiency and unsustainable testing caused by manual sampling and manual analysis in the past. It can realize the automatic upload of logs, measurement data, power environment status, system operation status and other data of the whole process of water quality monitoring, and can accept remote commands for debugging analysis, measurement analysis, etc.

(4) Networked medical equipment

Networked medical equipment is installed at both ends of doctors and patients. Signals can be transmitted via the Internet to achieve "mutual listening" and "mutual viewing" between doctors and patients. A basic remote consultation system is formed like this. Generally, telemedicine uses networked medical equipment to exchange information between doctors and patients far away to collect patient-related information. On the basis of collecting patient information, doctors can make corresponding diagnoses, propose medical plans, or even directly perform remote surgery.

(5) Application of network measurement and control technology in industrial field

Modern measurement and control technology is one of the core technologies of modern industry. The measurement and control systems are an important part of the production and processing equipment. In the production process, the automated control system and measurement equipment ensure the product quality.

(6) Application of networked measurement technology in agricultural production

In agricultural production, in order to improve the yield and quality of crops, it is necessary to monitor and adjust the agricultural production process. For example, it is necessary to analyze the moisture, pH, and seed health of the soil. Through the

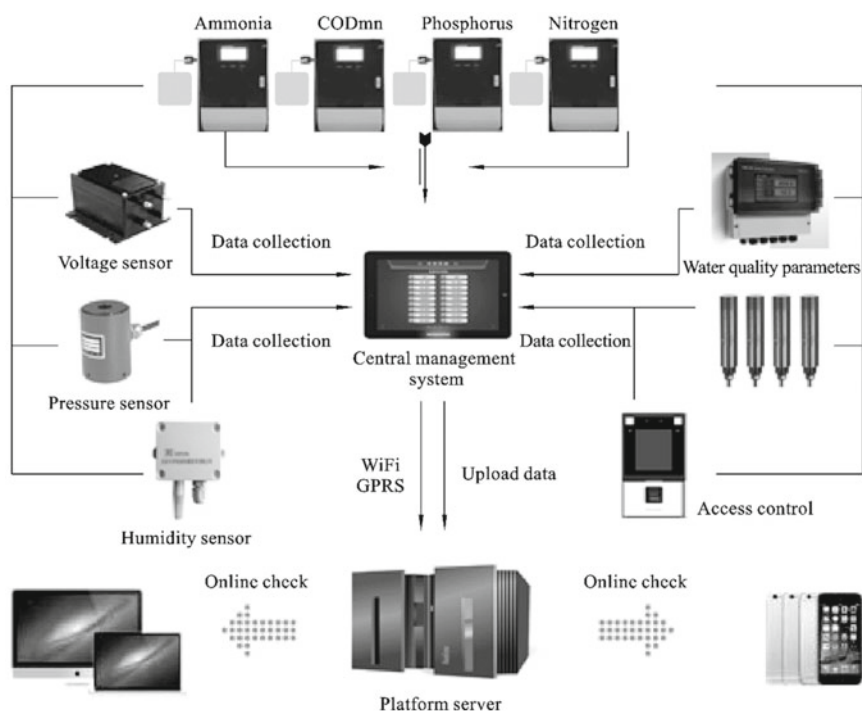


Fig. 10.10 Networked water quality monitoring system (image courtesy of Hasky environmental protection Inc.)

network measurement and control technology, these parameters can be monitored in real time with sensors, and the data can be transmitted to the farm management personnel through the network. Thus, the growth of crops can be known in time. The system can also analyze whether the crops will be affected by harsh environments such as drought and high temperature.

(7) Application of networked measurement technology in transportation

Networked measurement and control technology can be widely used in the monitoring of important facilities in the transportation field (such as pipelines, bridges, tunnels, railways, airports, highways, etc.) and transportation tools (such as airplanes, trains, automobiles, ships, etc.). Networked measurement and control can be used to effectively improve the traffic efficiency of intersections. It records traffic flow information in various directions through image sensors (cameras), and transmits this information to relevant personnel in the traffic control department through the network. The traffic flow conditions in different directions can be analyzed, and the red and green light durations can be adjusted in time to improve the traffic efficiency of congested road sections.

10.3.3 Internet of Things (IoT)

Internet technology has long enabled the interconnection of computers around the world, but the terminals of the Internet are limited to devices such as computers, tablets, and mobile phones. The concept of the Internet of Things is proposed on the concept of the Internet, and its terminals are embedded systems containing sensors, such as wearable devices, virtual reality systems, intelligent monitoring systems, and remote control systems. It extends the concept of the Internet to realize interconnection between things and things, people and things. The concept is shown in Fig. 10.11.

1 The system framework of the Internet of Things

The Internet of Things is a comprehensive, integrated and innovative technology system. According to the principles of information generation, transmission, processing, and application, the Internet of Things can be divided into a perception layer, a network layer, and an application layer. Figure 10.12 shows the three-layer structure of the Internet of Things.

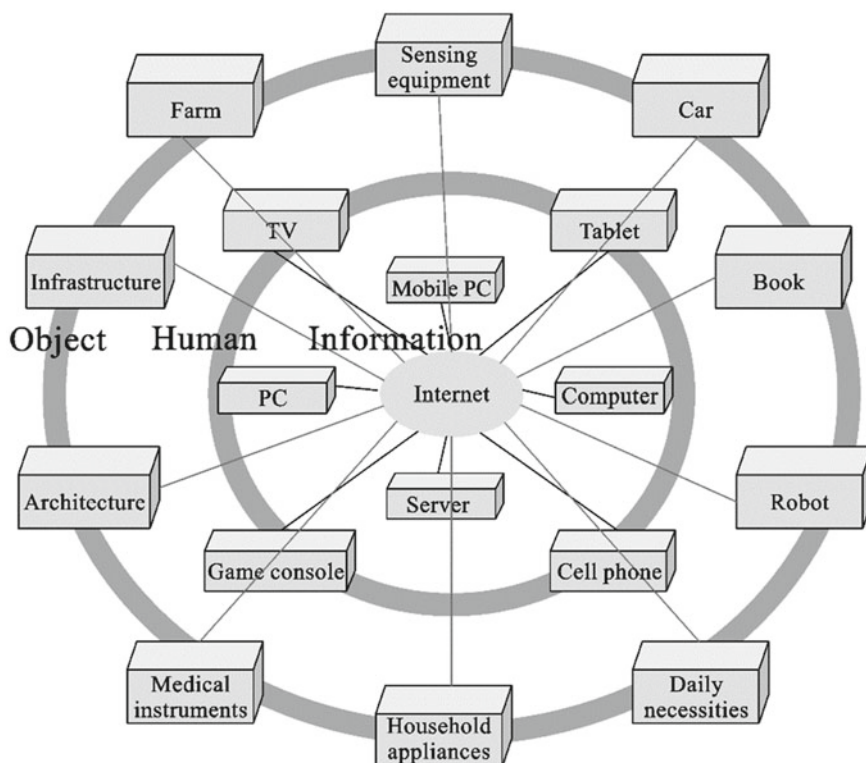


Fig. 10.11 Diagram of IoT

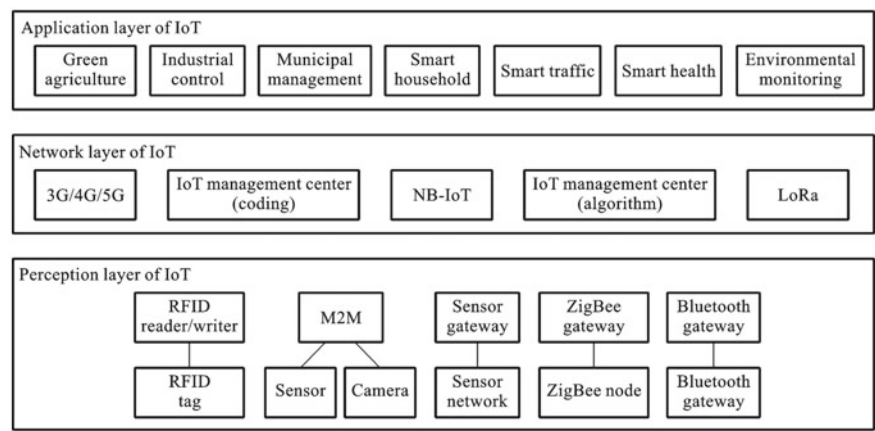


Fig. 10.12 The three-layer structure of the Internet of Things

(1) Perception layer

The perception layer is the lowest layer of the Internet of Things, a bridge that realizes the transformation from the physical world to the digital world, and is responsible for the generation of information. The perception layer obtains environmental information such as temperature, humidity, and light intensity in the physical world through various sensors, and stores them as digital signals to provide support for the dissemination and sharing of network information.

(2) Network layer

The role of the network layer is to transmit, exchange and integrate the information obtained by the perception layer, and its core is to realize the interconnection of information between different devices on the perception layer through the Internet. The interconnection of the perception layer information is mainly realized through two steps: (1) the gateway realizes the connection between the LAN and the WAN; (2) the TCP or UDP/IP realizes the spread of information in the WAN. The network technologies used mainly include: WiFi, Ethernet, Bluetooth, NFC, ZigBee and mobile network communications. WiFi is the most common wireless transmission technology, suitable for smart homes, smart offices and other occasions; Ethernet is suitable for cameras, alarms and other fixed devices that do not need to be moved frequently; Bluetooth is a short-distance wireless transmission method, suitable for low-frequency transmission such as earphones and low-power consumption equipment; ZigBee has low power consumption and multi-node processing capabilities, and is generally used for personal electronic product interconnection, industrial equipment control and other fields; mobile communication network has the advantage of wide coverage, but high power consumption, which needs to be used to facilitate device charging occasion.

(3) Application layer

The purpose of the application layer is to complete user-specified services, which is the ultimate goal of the Internet of Things. The application layer realizes intelligent perception recognition, positioning, traceability monitoring and management by analyzing and processing the information and data of the perception layer. For example, in a smart home system, the user can check the temperature information obtained by the temperature sensor in the room through the mobile phone on the way home, and turn on the air conditioner in advance before s/he reaches home.

2. Key technologies of the Internet of Things

(1) Sensing technology

Sensors are generally composed of components that are sensitive to a certain parameter. Their role in the Internet of Things is similar to that of human sense organs, which are used to sense and collect information in the environment, such as temperature, humidity, pressure, size, composition, and so on.

(2) Network communication technology

The realization of the Internet of Things requires the transmission and fusion of a large number of sensor data, so network communication technology plays a vital role. Network communication technologies include various wired and wireless transmission technologies and gateway technologies. Common wireless communication technologies include WiFi, Bluetooth, NFC, ZigBee and mobile network communication (2G/3G/4G/5G). With the development of 5G technology, the information transmission rate has greatly increased, and the transmission of big data from sensors has become easier, and the Internet of Things technology is expected to have a wider range of applications.

(3) Data analysis and processing technology

The Internet of Things technology is based on sensors. It needs to analyze and process the massive amounts of information collected by sensors, dig out valuable rules and conclusions, and provide the conclusions to decision-making users in the form of charts and other forms. In recent years, the rapid development of artificial intelligence (AI) technology and deep learning has provided good support for big data processing, which is expected to replace manual decision-making and control in the future.

3. Application of the Internet of Things

The Internet of Things has a wide range of applications, involving all aspects of life, such as intelligent transportation, environmental protection, public safety, smart home, industrial monitoring, agricultural production, food traceability and many other fields. The arrival of the 5G era provides high-quality conditions for the Internet of Things. With the advantages of high bandwidth and low latency, 5G provides a bearer for new services and new applications of the Internet of Things. Service applications that were once restricted by traditional mobile communications can be realized under 5G. With the help of 5G, the Internet of Things has gradually gained large-scale verification in the fields of smart cities, transportation, environmental protection, medical care, security, and electricity.

Exercise

1. What are the components of the computerized measurement system?
2. What is a virtual instrument? What are the characteristics of virtual instruments compared with traditional instruments?
3. What are the types of hardware configuration of virtual instruments? What is its core technology?
4. Try to design a spectrum analyzer by LabVIEW or DRVI software development platform.
5. Try to list the applications of virtual instruments in engineering.
6. Briefly describe the definition, characteristics and working modes of networked measurement instruments.
7. What are the key technologies of the Internet of Things?